LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Masterarbeit

im Studiengang Computerlinguistik

an der Ludwig-Maximilians-Universität München

Fakultät für Sprach- und Literaturwissenschaften

Few Shot Classification For Hate Speech Detection

vorgelegt von
Evgenii Kaurov

Betreuer:	Amir Hossein Kargaran
Prüfer:	Prof. Dr. Hinrich Schütze
Bearbeitungszeitraum:	13. März - 31. Juli 2023

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig angefertigt, alle Zitate als solche kenntlich gemacht sowie alle benutzten Quellen und Hilfsmittel angegeben habe.

München, den 31. Juli 2023

.....

Evgenii Kaurov

Abstract

In recent years, social media platforms have surged in user activity, becoming vital channels for social interaction, including governmental organizations and public figures. Safeguarding users from harmful content, like hate speech, has become paramount. Detecting and mitigating harmful language are critical for fostering a safer and inclusive online environment.

And yet, this task remains a significant challenge due to the diverse and sometimes vague semantic nature of harmful language, posing a considerable obstacle for machine recognition. Nevertheless, recent advances in machine learning, particularly transformer-based models like BERT, offer promising results. These models' capacity to learn from vast data enables them to capture nuanced patterns associated with hate speech. However, BERT-based models require specific labels during training, and adding new labels often involves discarding the classification head and retraining it from scratch, resulting in the loss of previously learned information.

A potential solution to this problem could be the Task-Aware Representation of Sentences (TARS) method, proposed by Halder et al., which proposes a new tactic of reducing the multilabel classification problem to a binary one while retaining information about the labels. TARS achieves this by essentially merging the labels with the sentences, allowing the model to learn more effectively from the available labeled data and retain previously acquired knowledge during training.

The main goal of our research is to investigate the efficiency of the TARS method in the context of hate speech detection. In our experiments, we employ cross-lingual language models XLM-RoBERTa-base and Glot500-base, subjecting them to consecutive and simultaneous training using diverse hate speech datasets. Our findings reveal the remarkable zero-shot and knowledge retention capabilities of the TARS method. Moreover, we analyze the impact of different languages on model performance, gaining valuable insights into the dynamics of multilingual hate speech detection.

The high performance achieved through simultaneous training, combined with the potential to adapt TARS approach's for closer collaborative work, highlights its potential as a powerful tool for future hate speech detection systems. This research advances the understanding of cross-lingual models and paves the way for further advancements in hate speech detection applications, fostering inclusive solutions to combat hate speech across languages.

Zusammenfassung

In den letzten Jahren verzeichneten soziale Medien eine immense Zunahme an Nutzeraktivität und wurden zu wichtigen Kanälen für soziale Interaktionen, einschließlich Regierungsorganisationen und öffentlicher Persönlichkeiten. Den Schutz der Nutzer vor schädlichen Inhalten wie Hate Speech zu gewährleisten, ist von größter Bedeutung. Die Erkennung und Eindämmung der Hassrede sind entscheidend für die Entwicklung einer sichereren und inklusiveren Online-Umgebung.

Trotzdem bleibt diese Aufgabe eine bedeutende Herausforderung aufgrund der vielfältigen und manchmal vagen semantischen Natur der Hassrede. Dennoch bieten aktuelle Fortschritte im maschinellen Lernen, insbesondere transformerbasierte Modelle wie BERT, vielversprechende Ergebnisse. Die Fähigkeit dieser Modelle, aus umfangreichen Daten zu lernen, ermöglicht es ihnen, feinere Muster der menschlichen Sprache zu erfassen. Allerdings erfordern BERT-basierte Modelle während des Trainings spezifische Labels, und das Hinzufügen neuer Labels erfordert oft das Verwerfen des Classification-Layers und ein erneutes Training von Grund auf, was zu einem Verlust zuvor gelernter Informationen führen kann.

Eine mögliche Lösung für dieses Problem könnte die "Task-Aware Representation of Sentences" (TARS) Methode sein, die von Halder et al. vorgeschlagen wurde. Sie schlägt einen neuen Ansatz vor, das Multilabel-Klassifikationsproblem auf ein binäres Problem zu reduzieren und gleichzeitig Informationen über die Labels beizubehalten. TARS erreicht dies hauptsächlich durch die praktische Fusion der Labels mit den Sätzen. So kann das Modell die in den Labels enthaltene Informationen trotz Reduktion bis auf die binäre Klassifikation behalten. Bei diesem Ansatz entfällt auch die Notwendigkeit, beim Hinzufügen neuer Labels das Classification-Layer zu verwerfen, und dadurch zuvor erlernte Informationen zu behalten.

Das Hauptziel unserer Forschung ist es, die Effizienz der TARS-Methode im Kontext der Hate Speech-Erkennung zu untersuchen. In unseren Experimenten verwenden wir die multilingualen Modelle XLM-RoBERTa-base und Glot500-base und unterziehen sie aufeinanderfolgendem und simultanem Training mit verschiedenen Hate Speech-Datensätzen. Unsere Ergebnisse zeigen bemerkenswerte Zero-Shot- und Knowledge-Retention-Fähigkeiten der TARS-Methode. Darüber hinaus untersuchen wir, wie sich das Modell verhält, wenn es mit mehreren Datensätzen in unterschiedlichen Sprachen trainiert wird, und gewinnen wertvolle Einblicke in die Dynamik der multilingualen Hate-Speech-Erkennung.

Die hohe Leistungsfähigkeit durch simultanes Training in Kombination mit der Möglichkeit, die TARS-Methode für eine engere Zusammenarbeit anzupassen, unterstreicht ihr Potenzial als leistungsstarkes Werkzeug für zukünftige Hate Speech-Erkennungssysteme. Diese Forschung fördert das Verständnis von mehrsprachigen Modellen und ebnet den Weg für weitere Fortschritte in der Hate Speech-Erkennung, um umfassende Lösungen im Kampf gegen Hassrede in verschiedenen Sprachen zu ermöglichen.

Contents

Abstract

Zusammenfassung **I**

1 Introduction **3**
1.1 Research Objective 4
1.2 Research Questions 5
1.3 Contributions of the Work 5
1.4 Organization of the Thesis 6

2 Theoretical Background **7**
2.1 Transformers: A Paradigm Shift in Natural Language Processing 7
2.2 HateBERT: Fine-tuning BERT for Abusive Language Detection in English 8
2.3 XLM and XLM-R: Multilingual Transformers 9
2.4 TARS: Task-Aware Representation of Sentences 9

3 Related Work **11**
3.1 Personal Attacks Seen at Scale 11
3.2 Continuous Hate Speech Measurement 12
3.3 Detecting Toxic Language in Brazilian Portuguese Social Media 13
3.4 Detecting Hate Speech and Offensive Language 14
3.5 GermEval 2018 Shared Task on Offensive Language Identification 15

4 Preparing Data **17**
4.1 Dataset Description 17
4.2 Data Processing 18
4.3 Criteria for Category Assignment 19

5 Settings and Experiments **21**
5.1 Settings 21
5.1.1 Models 21
5.1.2 Hyperparameters 22
5.1.3 Metrics 22
5.2 Experiments 22
5.2.1 Consecutive Training: Round 1, XLM-RoBERTa 22
5.2.2 Consecutive Training: Round 1, XLM-RoBERTa with Frozen Layers 23
5.2.3 Consecutive Training: Round 1, Glot500 23
5.2.4 Consecutive Training: Round 2, XLM-RoBERTa 23
5.2.5 Simultaneous Training: XLM-RoBERTa 24
5.2.6 Few-Shot (k-Shot) Learning with XLM-RoBERTa 24

6 Result Analysis **25**
6.0.1 Consecutive Training: Round 1, XLM-RoBERTa 25
6.0.2 Consecutive Training: Round 1, XLM-RoBERTa with Frozen Layers 26
6.0.3 Consecutive Training: Round 1, Glot500 27
6.0.4 Consecutive Training: Round 2, XLM-RoBERTa 27
6.0.5 Simultaneous Training: XLM-RoBERTa 28
6.0.6 Few-Shot (k-Shot) Learning with XLM-RoBERTa 29

7 Conclusion **31**
7.1 Knowledge Retention and Zero-Shot Capabilities 31
7.2 Impact of different languages 31
7.3 Simultaneous training 31
7.4 Few-shot (k-shot) learning 31
7.5 summary 32

8 Future work **33**

Bibliography **35**

List of Figures **37**

List of Tables **39**

1 Introduction

In recent years, social media platforms have experienced an unprecedented surge in user activity, transforming from simple communication tools into vital channels for social interaction. The widespread adoption of social media has extended beyond personal use, with governmental organizations, politicians, and public figures now utilizing these platforms as essential means of communication with the general public. As social media platforms have gained immense prominence, it has become crucial to protect users from harmful content, including hate speech. The detection and mitigation of harmful language emerged as critical tasks in ensuring users' safety and fostering a more inclusive online environment.

However, the detection of hate speech remains a significant challenge, mainly due to the inherent difficulty in defining the border between harmful language and acceptable language, making it even harder to teach machines to recognize it [3].

And yet despite these difficulties, current advances in machine learning have shown promising results in addressing this problem. The rise of transformer-based big language models, such as BERT [4], has revolutionized the field of natural language processing. One of the main features of BERT and similar transformer-based models is their ability to learn from large amounts of data, enabling them to capture nuanced patterns and features associated with hate speech and inappropriate content. The fine-tuning process allows the model to adapt to the specific task of hate speech detection, leading to improved precision and accuracy.

However, despite the potential of transformer-based models, they still require substantial amounts of labeled data for effective training, and the deficit of such data is one of the challenges faced in hate speech identification tasks [5].

But despite all the challenges there exist specialized models designed to address this issue. These models focus on specific labels or categories, such as HateBERT [1], which is a BERT-based model specifically designed for binary hate speech classification. HateBERT aims to differentiate between hateful and non-hateful content, providing a binary classification for hate speech detection. By fine-tuning BERT on hate speech datasets, HateBERT has shown promising performance in accurately identifying hateful language.

While HateBERT and simillar models have demonstrated effectiveness in identifying hate speech for specific labels, a significant challenge arises when incorporating new labels into these models. The standard approach involves cutting off the classification head and retraining the model from scratch with the new label set. However, this approach presents several challenges and limitations.

Firstly, cutting off the classification head and retraining the model from scratch results in the loss of previously learned information. The previously trained model may have captured valuable features and representations that are relevant to the new labels. By discarding the existing knowledge, the model essentially starts the learning process anew, potentially wasting the computational resources and time invested in the previous training.

Furthermore, retraining the model from scratch can be computationally expensive and time-consuming. Fine-tuning BERT models on large datasets is a resource-intensive task, requiring substantial computational power and extended training durations. When new labels need to be incorporated, this process needs to be repeated, adding further burden to the training pipeline and increasing the time and resource costs.

Recognizing the challenges posed by the inflexibility of traditional models when incorporating new labels, researchers have made notable efforts to find solutions. One such solution is the TARS (Task-Aware Representation of Sentences) method [6].

The TARS method introduces a flexible and knowledge-accumulating approach to generic

text classification. The authors propose transforming the multiclass classification problem into a series of binary classification problems, where each label is merged with the corresponding sentences. This novel formulation allows for the addition of new data without losing the information gained during previous training runs.

By merging labels with sentences, the TARS framework enables the model to learn not only from the textual content but also from the labels themselves. This approach facilitates knowledge accumulation, as the model continues to learn and adapt with the addition of new labeled data, building upon the existing knowledge and expanding its understanding of the classification task.

One particularly compelling aspect of the TARS approach is its ability to demonstrate surprising few and zero-shot learning capabilities. Few-shot learning refers to the model's ability to generalize and make accurate predictions with only a small number of labeled examples per class, while zero-shot learning allows the model to infer the classification of unseen labels without any labeled examples. This capability is particularly valuable in the field of hate speech identification, as new and emerging forms of abusive language can be rapidly detected and classified with minimal labeled data.

The TARS framework, with its flexible and knowledge-accumulating nature, presents a promising solution to the challenges of incorporating new labels without discarding previously learned information. This adaptability and capacity for continuous model improvement pave the way for effective hate speech identification and foster a more inclusive online environment. Moreover, the flexibility offered by the TARS method opens up possibilities for close community-wide cooperation, where different researchers can readily contribute to the model's progress by providing new data or even continuing the model training themselves. Such a collaborative approach harnesses collective knowledge and diverse datasets from various teams, creating a synergistic effort towards enhancing hate speech detection systems and promoting online safety and respect across different languages and cultural contexts.

1.1 Research Objective

The primary objectives of our research were as follows:

1. **Investigate the general performance of the TARS approach on semantically diverse and oftentimes unbalanced hate speech corpora:** We aimed to evaluate the effectiveness of the TARS framework in handling the challenges posed by hate speech datasets, which are known for their semantic diversity and imbalanced class distributions. By applying the TARS approach to hate speech detection, we sought to assess its ability to handle the nuances and complexities of hate speech data, considering the varying degrees of offensive and abusive language present in such corpora.

2. **Investigate the few-shot and zero-shot learning capabilities of the TARS approach:** We aimed to explore the TARS framework's capacity to generalize and make accurate predictions with limited labeled data per class (few-shot learning) and to infer the classification of previously unseen labels (zero-shot learning). By examining the few-shot and zero-shot learning abilities of the TARS approach in the context of hate speech identification, we aimed to assess its potential for rapid adaptation to emerging forms of abusive language and its capability to detect and classify new labels with minimal labeled data.

3. **Investigate the capabilities of knowledge retention within the TARS framework:** We aimed to explore how well the TARS approach retains previously learned information when incorporating new labels and additional labeled data. By investigating the knowledge retention capabilities of the TARS framework, we sought to

determine whether the model could accumulate knowledge from previous training runs and effectively utilize it when confronted with new data. This aspect was particularly relevant in the context of hate speech detection, where the dynamic nature of abusive language necessitates continuous model updates and adaptations.

4. **Investigate the performance of the TARS method with datasets in different languages:** We aimed to assess the performance of the TARS approach when working with hate speech datasets in different languages. By experimenting with diverse datasets in English, German and Portuguese, we sought to gain insights into the TARS method's ability to handle multilingual data and analyze how different languages may impact the model's performance in hate speech detection tasks. This objective was particularly interesting, as the original paper did not evaluate the method's performance in the context of different languages.

By addressing these research objectives, we aimed to contribute to the understanding of the TARS approach's applicability and effectiveness in hate speech detection, highlighting its potential to handle semantically diverse data, adapt to emerging forms of abusive language, and retain previously acquired knowledge.

1.2 Research Questions

In this chapter, we outline the research questions that guided our investigation and delve into the answers we obtained through our experiments. The following were the main questions we sought to address:

1. **What will be the overall TARS performance on semantically challenging data such as hate speech corpora?** We aimed to evaluate the effectiveness of the TARS approach in handling the complexities and semantic diversity present in hate speech datasets. By assessing its performance on such challenging data, we sought to gain insights into its suitability for hate speech detection tasks.

2. **How good are the zero-shot capabilities of the TARS method?** Zero-shot learning is crucial in hate speech detection, as it allows the model to classify previously unseen labels without any labeled training data. We aimed to explore the TARS framework's ability to generalize and make accurate predictions in zero-shot scenarios.

3. **What influence does the sample size have in few-shot scenarios?** Few-shot learning is valuable when labeled data is scarce. We aimed to investigate how the TARS method performs with limited labeled data per class and determine its adaptability in few-shot conditions.

4. **Which training mode is the most efficient?** We sought to compare the performance of TARS under consecutive training and simultaneous training modes. Understanding the strengths and weaknesses of each training mode would provide valuable insights into the most effective approach for hate speech detection.

5. **What is the TARS performance when we have data in different languages?** The original TARS paper focused on data in English only. We aimed to assess how well the TARS method performs when dealing with hate speech datasets in different languages.

1.3 Contributions of the Work

1. We conducted a comprehensive evaluation of the TARS method using three models on five hate speech datasets in English, Portuguese, and German. These datasets

varied significantly in size and label distribution. By comparing different training modes, we identified the most efficient approach, which achieved an outstanding average F1 score of **0.86**.

2. Our experiments reaffirmed the remarkable zero-shot learning capabilities of the TARS method, validating the promising results reported in the original TARS paper.

3. We made significant contributions by investigating the dynamics of hate speech detection when datasets are available in different languages. This aspect had not been extensively studied in the original TARS research.

4. Through rigorous few-shot learning experiments, we observed that the TARS method's performance remained consistent across different sample sizes less than 100. The model displayed adaptability and effectiveness even with limited labeled data.

1.4 Organization of the Thesis

- **Chapter 1:** Provides an overview of the research objectives, research questions, and research contributions. Outlines the structure and content of the thesis.
- **Chapter 2:** Discusses important works that serve as the theoretical background for our research, with a focus on Section 2.4 on the TARS method.
- **Chapter 3:** Introduces related research that provided the datasets used in our work. Provides details about the datasets employed in our experiments.
- **Chapter 4:** Thoroughly describes the data preprocessing steps we performed before feeding it into the model, including data cleaning and transformation processes.
- **Chapter 5:** Describes the experimental settings, including the selection of models and hyperparameters. Outlines the procedural steps conducted during our experiments.
- **Chapter 6:** Extensively explains the results obtained from the experiments conducted in Chapter 5. Presents the performance and analysis of the TARS method on different datasets.
- **Chapter 7:** Contains our conclusions where we summarize our findings and discuss the implications of our research. Highlights the contributions and limitations of our work.
- **Chapter 8:** Describes potential interesting topics for future research and advancements in the field of hate speech detection using the TARS method.

2 Theoretical Background

In the section titled "Theoretical Background," we explore fundamental concepts and methodologies that underpin the research. We discuss transformers, including the groundbreaking BERT model. Additionally, we highlight the potential of retraining BERT for specific tasks and explore multilingual transformers like XLM and XLM-R. Finally, we introduce the TARS framework, addressing challenges in text classification tasks.

2.1 Transformers: A Paradigm Shift in Natural Language Processing

Transformers, introduced in the groundbreaking paper "Attention is All You Need" by Vaswani et al. [12], revolutionized sequence modeling tasks by surpassing the limitations of traditional recurrent neural networks (RNNs). The Transformer model leverages self-attention mechanisms to capture global dependencies and contextual information in text, enabling more effective language modeling.

One of the most influential applications of the transformer architecture is BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al [4]. BERT was motivated by the need to overcome limitations in previous contextual language models. Unlike uni-directional models that read text in one direction, BERT captures bidirectional contextual information by pre-training on vast amounts of unlabeled data using masked language modeling and next sentence prediction tasks. This process enables BERT to learn deep contextual understanding, making it highly effective in a wide range of natural language understanding tasks.

BERT's architecture consists of multiple encoder layers, each equipped with self-attention mechanisms and feed-forward layers, enabling the model to encode complex linguistic patterns and contextual relationships within the input text. The power of BERT lies not only in its architecture but also in its two-step pre-training and fine-tuning process. During pre-training, BERT learns contextual representations by predicting masked words in a sentence and determining relationships between two sentences. In the fine-tuning stage, BERT's pre-trained weights are adapted to specific downstream tasks, such as text classification or named entity recognition, making it highly versatile across various NLP domains.

Since the introduction of BERT, the NLP research community has witnessed a rapid proliferation of transformer-based models, each designed to address specific challenges in natural language understanding. For instance, the GPT (Generative Pre-trained Transformer) series, including GPT-1, GPT-2, and GPT-3, developed by OpenAI [10][11][12], focused on language generation and demonstrated remarkable performance in tasks like text completion and story generation.

Moreover, XLM (Cross-lingual Language Model), a multilingual transformer model introduced by Conneau et al. (2019) [8], has gained significant attention for its ability to handle multiple languages. Through learning to predict masked words and translation language modeling, XLM leverages large-scale parallel data from various languages during training, acquiring a deep understanding of multilingual contexts and facilitating knowledge transfer across languages.

Another notable advancement is XLM-R (XLM-RoBERTa), introduced by Conneau et al. (2019) [2], which builds upon XLM with improvements in training methodology, pre-training data size, and model size. XLM-R achieved state-of-the-art results in various NLP tasks across different languages, showcasing its effectiveness and versatility as a

multilingual transformer model.

The transformative impact of transformers extends beyond BERT and XLM. Researchers continue to explore different variations and adaptations of transformer-based models to achieve state-of-the-art results in various NLP domains. The success of transformers in natural language processing has ushered in a new era of language understanding, empowering models with the ability to capture deep contextual information and unlock the potential for a wide range of natural language tasks.

2.2 HateBERT: Fine-tuning BERT for Abusive Language Detection in English

In the conducted experiments with HateBERT [1], the researchers assessed its effectiveness in detecting abusive language phenomena on three English datasets: OffensEval 2019, AbusEval, and HatEval. Each dataset represented different aspects of offensive language, abusive language, and hate speech, allowing for a comprehensive evaluation of HateBERT's performance across varied language phenomena.

In the first dataset, OffensEval 2019 [16], containing 14,100 tweets annotated for offensive language, HateBERT demonstrated remarkable results. It outperformed the generic BERT model with a macro F1 score of 0.809±0.008, compared to BERT's 0.803±0.006. Notably, HateBERT achieved an F1 score of 0.723±0.012 for the positive class, whereas BERT's corresponding score was slightly lower at 0.715±0.009. These findings showcased the superior performance of HateBERT in detecting offensive language, highlighting its potential in handling non-acceptable language and targeted offenses.

Next, the researchers evaluated HateBERT's performance on the AbusEval dataset, which was an extension of OffensEval 2019 with an additional layer of abusive language annotation. AbusEval focused on hurtful language used to insult or offend others based on personal qualities, appearance, social status, opinions, statements, or actions. HateBERT achieved an impressive macro F1 score of 0.765±0.006, surpassing BERT's score of 0.727±0.008. In terms of the positive class F1 score, HateBERT achieved 0.623±0.010, while BERT scored 0.552±0.012. These results demonstrated HateBERT's enhanced capability in detecting abusive language, particularly in situations where the offensive language takes on a more targeted and harmful nature.

Furthermore, the researchers evaluated HateBERT on the HatEval dataset, which contained 13,000 tweets annotated for hate speech against migrants and women. Hate speech was defined as communication that disparages a person or group based on characteristics like race, color, ethnicity, gender, sexual orientation, nationality, religion, or other factors. HateBERT exhibited a macro F1 score of 0.516±0.007, while BERT's score was 0.480±0.008. Moreover, HateBERT achieved a positive class F1 score of 0.645±0.001, whereas BERT scored 0.633±0.002. These outcomes revealed HateBERT's effectiveness in detecting hate speech directed at specific targets like migrants and women, showcasing its capacity to identify harmful language used against particular groups.

The in-dataset evaluation consistently demonstrated HateBERT's superiority over the generic BERT model, with higher macro F1 scores and positive class F1 scores on all three datasets. Notably, HateBERT's performance even outperformed the state-of-the-art model on the AbusEval dataset, further substantiating its effectiveness in detecting abusive language phenomena. The success of HateBERT in these experiments highlighted the potential of retraining transformer-based models like BERT for specific tasks, enabling their adaptation to diverse language varieties and domains, even when faced with challenges such as limited and non-standard datasets.

2.3 XLM and XLM-R: Multilingual Transformers

XLM (Cross-lingual Language Model) and its advanced version XLM-R (XLM-RoBERTa) have emerged as significant developments in the field of natural language processing (NLP) due to their ability to handle multiple languages and achieve state-of-the-art performance in various tasks. XLM, a multilingual transformer model, gains its multilinguality by leveraging large-scale parallel data from diverse languages during training. Like BERT, it employs masked language modeling, but it also incorporates "translation language modeling" to align word and sentence representations across languages. This enables XLM to grasp shared semantic and syntactic structures, making it highly versatile across different language varieties [2].

XLM-R, an extension of XLM, represents a significant advancement in multilingual transformer models. It follows a similar architecture but incorporates improvements in training methodology, pre-training data size, and model size. The authors of XLM-R expanded the training data by including additional web crawled data and extensively pre-trained the model on monolingual data from 100 languages. This enhanced training approach contributes to XLM-R's outstanding performance on various NLP tasks across different languages.

While multilingual models like XLM-R offer numerous benefits, there is a trade-off associated with multilinguality, often referred to as the "curse of multilinguality." As the number of languages and data diversity increases, the performance of multilingual models may saturate or decline compared to monolingual models. However, researchers have found that increasing model size can help mitigate the impact of this curse and maintain performance levels.

Studies have demonstrated the state-of-the-art performance of XLM-R on diverse NLP tasks across various languages. For instance, Wang et al. [13] achieved impressive accuracy results with XLM-R on the SuperGLUE benchmark dataset.

The fine-tuning process for XLM-R is akin to BERT. Researchers add a classification layer on top of the pre-trained XLM-R model and fine-tune it on labeled data for specific tasks. While this process facilitates model adaptation, adding new labels may raise challenges and the potential loss of previously acquired knowledge.

In the context of NLP, XLM and XLM-R have significantly advanced multilingual transformer models, enabling them to comprehend diverse linguistic patterns and excel in various NLP tasks across different languages. Their success showcases the transformative impact of transformer-based models in empowering language understanding and addressing the challenges posed by multilingual contexts.

2.4 TARS: Task-Aware Representation of Sentences

The TARS (Task-Aware Representation of Sentences) paper proposed by Halder et al. [1] presents a novel approach to address the challenges of information loss and inflexibility in text classification tasks. The authors aimed to create a model that could incorporate new labels without discarding previously learned information.

The main idea behind TARS is to transform a multi-label classification problem into a binary classification problem while retaining the information about specific labels. They introduced a universal binary text classification formulation, where each label is treated as an individual binary classification task. The model takes a tuple of both the text input and the class label as input and predicts whether there is a match between the label and the text.

By adopting this approach, TARS provides a mechanism to both simplify the task by turning it into a binary classification problem and retain the information about specific labels. The architecture of the TARS model is designed to accommodate multiple tasks, allowing it to be used across various text classification problems. The results presented

in the TARS paper demonstrate its effectiveness. TARS outperforms the baselines in zero-shot classification, where it achieves considerably higher accuracy than the random baseline. It also shows stronger few-shot learning capabilities, adapting quickly to the target task and achieving higher accuracy scores with minimal training data, seemingly learning from the labels themselves. However, the advantage of TARS over the baselines diminishes as the models receive more training data, indicating that TARS is particularly beneficial in scenarios with limited training data. The effectiveness of transfer learning in TARS is also explored, considering semantically different datasets. The results reveal that TARS outperforms the baselines in transfer learning, even across tasks with varying semantic distance. However, in some cases where the tasks exhibit significantly different language and domain, a BERT-base model trained directly on target task data performs better than transfer learning approaches. Nonetheless, TARS still demonstrates its robustness in transfer learning, surpassing BERT-base in those scenarios.

The TARS approach addresses the challenges of information loss and inflexibility in text classification tasks. By transforming the multi-label classification problem into a binary classification problem, TARS retains information about specific labels while simplifying the task. It demonstrates promising performance in zero-shot and few-shot learning scenarios, showcasing its effectiveness in handling limited training data. The success of transfer learning in TARS is influenced by the semantic distance between tasks, highlighting its versatility and adaptability across different text classification domains.

3 Related Work

In this section, we present the related works that have significantly contributed to the field of hate speech detection and have produced the datasets utilized in our research. These works encompass studies focusing on hate speech identification, annotation reliability, and offensive language detection. Their main relevancy for our work is because we used the datasets produced during these studies.

3.1 Personal Attacks Seen at Scale

The research conducted by Wulczyn, Thain, and Dixon [15] focuses on studying personal attacks in online discussions, particularly on the English Wikipedia platform. The motivation behind the research stems from the prevalence of online harassment and the need to develop effective policies to identify and respond to such behavior. The researchers aim to analyze a large corpus of online comments to understand the nature of personal attacks and contribute to reducing toxic discussions.

To build their dataset, the researchers employed crowdsourcing, where they labeled a small fraction of the corpus by determining whether each comment constitutes a personal attack. This labeled data was used to train a machine learning classifier, primarily utilizing character-level n-grams as features. The researchers experimented with different labeling methods and features, validating and extending previous findings on the effectiveness of character-level n-grams for detecting abusive language in English.

The trained classifier was then used to annotate the entire corpus, acting as a surrogate for crowd-workers. The researchers developed an evaluation method to compare the performance of the classifier with that of human annotators, showing that their classifier is as good at generating labels as aggregating judgments from three crowd-workers. The annotated corpus, along with the trained classifier, was made publicly available to support further research and independent replication.

For further analysis, the researchers applied the classifier's annotations to quantify and explore personal attacks within different sub-groups of comments. They addressed questions regarding the impact of anonymous contributions, variations in attacks based on the quantity of a user's contributions, concentration of attacks among specific users, the occurrence of moderator actions, and patterns in the timing of personal attacks.

The evaluation of different model architectures and label types revealed that character n-grams outperformed word n-grams in detecting personal attacks. Furthermore, training the models using empirical distribution labels, representing the fraction of annotators who considered a comment an attack, consistently improved performance compared to majority class-based labels. The evaluation metrics, including area under the receiver operating characteristic curve (AUC) and Spearman rank correlation, were reported for each model architecture and label type.

After testing several model-feature type combinations the authors reported multi-layer perceptron with character-level n-grams as features being the best model. The provided the following metrics: Area under the receiver operating characteristic curve (AUC) of 96.59 and the Spearman coefficient 67.

3.2 Continuous Hate Speech Measurement

The research "Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application" by Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano [7] addresses the limitation of discrete variables in hate speech research, which may result in the loss of valuable information. To overcome this, they propose a methodology that combines many-facet Rasch measurement and supervised deep learning.

The authors addressed the limitation of discrete variables in hate speech research, which may result in the loss of valuable information. To overcome this, they proposed a methodology that combines many-facet Rasch measurement and supervised deep learning. Their approach aimed to construct continuous, interval variables for arbitrary variables by emulating physical measurement systems' continuous scales, like temperature. They applied the method to hate speech measurement, a complex linguistic phenomenon with significant social and political impacts.

By using crowdsourced labeling and a theoretical construct with 8 levels, they created a labeling instrument with 32-48 measurement levels to accurately distinguish between theorized construct levels and reduce measurement error. The research showcased the effectiveness of their method in measuring hate speech, providing a more nuanced understanding of hateful content in speech with multiple measurement levels and reduced measurement error. This approach has potential applications in various fields where human-generated variables require continuous measurement scales.

The authors collected comments from YouTube, Twitter, and Reddit to create their dataset. They used public APIs to download recent English comments that were not too short or too long. The collection took place between March and August 2019. To label the comments, crowdsourcing was employed, and they used a stratified sampling approach to ensure an even distribution of labeled comments across 8 levels of hate speech. The stratification was based on relevance estimates of containing identity group references and a hypothesis score for the level of hate speech. Certain bins, such as potential counterspeech or violent hate speech, were heavily oversampled to improve efficiency. The labeled data consisted of 40% from Reddit, 40% from Twitter, and 20% from YouTube. The comments were grouped and randomly allocated for ratings by reviewers to ensure coverage of the entire hate speech spectrum.

The authors used Rasch measurement theory to convert the ordinal ratings from human reviewers into a continuous, linear hate speech scale. They employed the faceted partial credit model, which extended the Rasch family of models to include judge-mediated assessments, treating the rater as an additional facet. This allowed the estimation of a rater "severity" parameter, which represents the survey interpretation bias. The faceted Rasch models consider within-rater consistency rather than inter-rater reliability as the primary quality metric for dataset labeling. The scaling was conducted using Facets software, and four scaling estimates were performed to improve the estimates' quality by removing low-quality raters and collapsing response options.

After scaling, they trained a deep learning algorithm to map raw text to the latent hate speech score. The approach involved predicting the responses to each survey item using a multitask architecture, analyzing each item as an ordinal outcome, and incorporating the rater's interpretation bias. Slurs in comments were also tagged for comparison with human annotators.

In another approach, they hypothesized that an improved architecture would predict human ratings on each survey item. The comment text was processed through a deep natural language algorithm, and fully connected layers learned to predict item responses based on the vector representation and rater bias. The predicted responses were transformed into the continuous hate speech score via IRT, providing direct supervision during the optimization process and eliminating the need for extensive hyperparameter optimization. They used a multitask architecture to predict the responses to each survey

item efficiently, improving generalization. They implemented the consistent rank logits (CORAL) method to handle the ordinal nature of the item ratings, allowing for unimodal probability distributions over the possible item ratings. After training the model, they transformed the item-level predictions using the partial credit IRT scaling procedure. To address downsides of the partial credit scaling, they employed plausible value sampling, generating more precise predictions by sampling possible item ratings for each comment based on the predicted probability distributions.

In their evaluation, they analyzed the network linkage across raters, original comments, and reference set comments, ensuring the generation of a single linked network with no disjoint subsets. The faceted partial credit model achieved high case reliability (0.94), and rater separation reliability was also 0.94, indicating accurate estimation of individual severity for the labelers. The item difficulties were consistent with their hypothesized values, validating their instrument's construct. The calibrated Wright Map showed the scale across comments, items, item steps, and raters.

Regarding deep learning, their best models utilized RoBERTa-Large for text representation, achieving the lowest root mean-squared error (RMSE) in direct prediction of the continuous hate score. The multitask networks transformed via IRT achieved comparable performance, with slightly lower mean absolute error (MAE) and the advantage of explainability. Ordinal modeling of the items did not show a significant benefit compared to categorical modeling. The best model achieved an RMSE of 1.078 and a correlation of 0.839 in the validation data.

3.3 Detecting Toxic Language in Brazilian Portuguese Social Media

The research "Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis" by João Augusto Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton [9] addresses the rising concern of hate speech and toxic comments on social media platforms. The authors propose the "ToLD-Br" dataset, a large-scale collection of Twitter posts in Brazilian Portuguese manually annotated into seven toxicity categories. BERT models fine-tuned on monolingual data achieve up to 0.76 macro-F1 score in binary classification, emphasizing the importance of language-specific datasets. Transfer learning and zero-shot learning have limited success, highlighting the need for targeted approaches. An error analysis reveals challenges in classifying categories with fewer examples, calling for more data in those areas. The ToLD-Br dataset contributes significantly to research in this area, but further investigation is required to develop models aware of different categories of toxicity.

The data collection involved using the GATE Cloud's Twitter Collector tool from July to August 2019. Two strategies were employed: one based on predefined toxic keywords and hashtags, and the other targeting tweets mentioning influential users. Over 10 million unique tweets were collected, from which 21K examples were randomly selected for annotation.

The annotation process started by choosing volunteers from the Federal University of Sao Carlos, aiming to balance annotation bias by considering demographic information. Each annotator labeled 1,500 tweets into categories such as LGBTQ+phobia, obscene, insult, racism, misogyny, and xenophobia. Each tweet was annotated by three different annotators.

The analysis of the annotations showed variations in agreement among the classes, with LGBTQ+phobia demonstrating the highest agreement and obscene and racism showing the lowest. Moreover, the classes obscene and insult displayed confusion among annotators, suggesting an intersection between these categories, which was confirmed by common words in their top ten frequent words.

The researchers propose ToLD-Br, a large-scale dataset for identifying toxic comments

in Brazilian Portuguese social media. The dataset comprises tweets manually annotated into seven categories. Monolingual BERT models achieved up to 0.76 macro-F1 score on toxic comment classification. They explored label aggregation strategies and used the least restrictive approach to identify offensive posts comprehensively. Evaluation metrics included precision, recall, F1-score, and macro-F1 for binary classification. An initial experiment with multi-label classification using BoW+AutoML was also performed.

In binary classification, the best model was M-BERT-BR, achieving a macro-F1 score of 0.76. It outperformed other models like BR-BERT (macro-F1 score of 0.75) and M-BERT(transfer) (macro-F1 score of 0.56). The worst performing model was M-BERT(zero-shot) with a macro-F1 score of 0.43.

For multilabel classification, the baseline model using BoW+AutoML achieved a Hamming loss of 0.08 and an average precision of 0.20. The BERT-based model (M-BERT-BR) performed slightly better with a Hamming loss of 0.07 and an average precision of 0.19. However, both models faced challenges in dealing with the imbalanced dataset and the limited number of positive examples for some toxic classes.

Additionally, the analysis of false negatives in the binary classification revealed that classes with fewer examples, such as racism, misogyny, xenophobia, and LGBTQ+phobia, were more likely to be misclassified as non-toxic. The performance of the models improved as the size of the training dataset increased, emphasizing the importance of having a large-scale dataset like ToLD-Br.

3.4 Detecting Hate Speech and Offensive Language

The paper by Davidson et al. [3] delves into the challenge of automated hate speech detection on social media, specifically addressing the difficulty in distinguishing hate speech from other forms of offensive language. The authors emphasize the limitations of lexical detection methods, which often suffer from low precision by classifying all messages containing specific terms as hate speech. Supervised learning approaches have also struggled to effectively differentiate between hate speech and offensive language.

To address this issue, the authors present their approach using a crowd-sourced hate speech lexicon to collect tweets containing hate speech keywords. The crowd-sourced data is then labeled into three categories: tweets containing hate speech, those with offensive language but not hate speech, and those with neither hate speech nor offensive language. This multi-class classification approach enables the differentiation between the various categories.

The paper defines hate speech as language that targets disadvantaged social groups in a potentially harmful manner, while offensive language includes terms that are highly offensive to certain groups but may not necessarily express hatred. The study uses machine learning techniques, including logistic regression with L2 regularization and linear SVMs, along with various linguistic and syntactic features, to build a hate speech detection model.

The results show that the hate speech detection model performs well, achieving high precision and recall for prevalent forms of hate speech, such as anti-black racism and homophobia. However, it is less reliable at detecting rarer types of hate speech that occur infrequently. The model is also biased towards classifying tweets as less hateful or offensive than human coders, which highlights the importance of considering context in hate speech detection.

The study emphasizes the importance of accurately distinguishing between hate speech and offensive language, as misclassification can lead to significant legal and moral implications. While lexical methods are effective in identifying offensive terms, they may not accurately detect hate speech. The paper suggests that finding sources of training data with hateful content that does not necessarily use specific keywords or offensive language could improve hate speech classification.

Moreover, the authors highlight the subjective nature of hate speech classification and

the potential biases in human and automated approaches. They suggest that future research should explore the social contexts and motivations behind hate speech usage and focus on understanding the individuals who engage in hate speech. Correcting biases in hate speech detection algorithms is crucial to ensure fairness and accuracy in combating hate speech and offensive language on social media platforms.

3.5 GermEval 2018 Shared Task on Offensive Language Identification

The paper titled "Overview of the GermEval 2018 Shared Task [14] on the Identification of Offensive Language," authored by Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer in 2018, presents an overview of the GermEval 2018 Shared Task, which focuses on the identification of offensive language in German tweets.

In the coarse-grained task, tweets were classified as either "OFFENSE" or "OTHER," while the fine-grained task involved four categories: "OTHER," "PROFANITY," "INSULT," and "ABUSE." "PROFANITY" used profane words without intending to insult, "INSULT" aimed to offend someone, and "ABUSE" represented the strongest form of abusive language.

The shared task received significant participation with 20 teams submitting 51 runs for the coarse-grained task and 25 runs for the fine-grained task. Evaluation metrics included precision, recall, F1-score, and macro-average scores for each task. The organizers provided an evaluation tool for assessing participant performance.

The data for the GermEval 2018 Shared Task was collected from Twitter, making it publicly available. Tweets were sampled from various users' timelines to ensure a diverse vocabulary of offensive content. Non-offensive tweets were also included to balance the dataset. The data collection imposed formal restrictions on the tweets, such as being written in German, containing at least five alphabetic tokens, and no URLs or retweets.

The data collection, consisting of 8,541 tweets, was manually annotated by native German speakers among the organizers. Tweets marked as incomprehensible or exempt were removed. The remaining tweets were annotated by one annotator. The class distribution showed non-offensive tweets were the majority, followed by abuse, insults, and finally, profane tweets.

The data is distributed in tab-separated value files, representing each tweet with its text, coarse-grained label, and fine-grained label.

In the GermEval 2018 Shared Task, 20 teams participated in Task 1, and 11 also took part in Task 2. Various approaches were used, detailed in the teams' dedicated system description papers.

The classifiers used in the GermEval 2018 Shared Task included familiar non-neural types such as SVM, logistic regression, Decision Trees, and Naive Bayes. Among the neural network classifiers, common architectures like CNN, LSTM, and GRU were used, often in combination.

The results of the shared task indicated that the top-performing system for the coarse-grained binary classification task achieved an F1-score of 0.77. This system, TUWienKBS, utilized traditional supervised learning methods. For the fine-grained 4-way classification task, the best-performing system was uhhLT with an F1-score of 0.53. This system made use of neural network classifiers. Overall, the task of identifying offensive language in German tweets remains challenging, with F1-scores ranging from 0.32 to 0.77.

4 Preparing Data

In this section, we describe the data processing steps undertaken to implement the TARS-method. The selected datasets were transformed into a specific format, with each sentence associated with multiple rows, corresponding to different labels. We tailored a relabeling algorithm for each dataset to assign appropriate category labels. The criteria for label assignment varied based on the dataset's original labeling scheme. The resulting datasets are now suitable for exploring hate speech detection using the TARS-method across diverse contexts.

4.1 Dataset Description

In our search for suitable labeled hate speech data, we aimed to find datasets that offered unique labels and nuanced categories, avoiding binary datasets with simplistic classifications like "hate speech" and "normal." This approach added complexity to the task, as we sought datasets with more comprehensive and diverse labeling schemes to ensure the inclusion of varied linguistic contexts and hate speech manifestations.

After conducting a thorough inspection of available resources, we chose five datasets that best met our predefined criteria. However, the scarcity of publicly available data meeting our specific criteria made the selection process more challenging. In Table 4.1 we present essential information about the datasets, including their sources, aliases used in our code, specific labels derived from each dataset, original lengths, and the languages in which the data was composed. This comprehensive overview enables a holistic understanding of the dataset characteristics and facilitates robust analysis within distinct linguistic contexts.

Research Title	**Alias**	**Labels**	**Original length**	**Lang.**
Ex Machina: Personal Attacks Seen at Scale [15]	wiki	identity_attack, insult, obscene, severe_toxicity, threat, toxicity.	223549	EN
Constructing interval variables via faceted Rasch measurement and multitask deep learning [7]	measuring_hs	respect, insult, humiliate, status, dehumanize, violence, genocide	135 556	EN
Toxic Language Detection in Social Media for Brazilian Portuguese [9]	ToLD	homophobia, obscene, insult, racism, misogyny, xenophobia	21 000	PT
Automated Hate Speech Detection and the Problem of Offensive Language [3]	online_hs_recog	hate_speech, offensive_language	24783	EN
Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. [14]	GermEval	insult, profanity, abuse	8407	DE

Table 4.1: Short description of the datasets used in this research.

4.2 Data Processing

To implement the TARS-method, we undertook a data transformation process to convert the datasets into a specific format, as illustrated in Table 4.2.

Text	Label
Label [SEP] sentence	0 or 1

Table 4.2: The end-format of the data we used in the experiments.

For this purpose, we applied a general relabeling algorithm, which was further customized for each individual dataset. The objective was to remove unnecessary columns, assign appropriate label names to each sentence, and assign corresponding label values.

Next, we illustrate the steps taken to convert the data from the wiki dataset to the desired format:

1. Table 4.3 shows the initial state: the sentences and their corresponding labels in string format. We refer to the original columns with information about categories as "label columns."

2. Since there are 6 labels in the dataset, we included 6 copies of each sentence in the processed version.

3. All 6 labels were attached to the string containing the sentence itself, separated by a generic token (to be replaced with a model-specific one during tokenization).

4. The label value in the row with the true label was set to 1, and all others were set to 0. The formatted version of this step is presented in Table 4.4.

Identity attack	Insult	Obscene	Threat	Toxicity	Severe toxicity	Text
1	1	0	0	1	0	Singing is one thing, but you also need to show that you're a smart person and not some illiterate redneck from the south.

Table 4.3: An example of a raw sentence taken from the wiki-dataset.

As we see in Table 4.4, each sentence was then duplicated to match the number of categories. Each duplicated sentence corresponds to a specific category column in the original format. We refer to these duplicated sentences as "label rows."

Text	Label
Identity attack [SEP] Singing is one thing, but ...	1
Insult [SEP] Singing is one thing, but you also ...	1
Obscene [SEP] Singing is one thing, but you also ...	0
Threat [SEP] Singing is one thing, but you also ...	0
Toxicity [SEP] Singing is one thing, but you also ...	1
Severe toxicity [SEP] Singing is one thing, but ...	0

Table 4.4: The formatted version of the sentence from Table 4.3.

It's important to note that some datasets originally had categories like "other," "neither," or "counter speech." However, we only considered categories with harmful language. Sentences in the normal language categories were not discarded but received 0 in the "label" column, making them purely negative examples.

Table 4.5 illustrates an example of a processed sentence from the category "other" in the GermEval dataset.

Text	Label
insult [SEP] @user tja, es können einen...	0
profanity [SEP] @user tja, es können einen...	0
abuse [SEP] @user tja, es können einen...	0

Table 4.5: An example of a formatted sentence from a non-harmful category from the GermEval dataset.

As a result of this processing, all datasets increased in size accordingly:

$$(\text{original length} - \text{number of too long sentences}) \times \text{number of categories}$$

Note: BERT-based models can only handle sequences up to 512 tokens in length. As we wanted to avoid truncation to prevent information loss, we manually deleted all rows with sentences longer than 512 symbols.

4.3 Criteria for Category Assignment

The datasets we had at our disposal were very diverse, each with its unique marking and categories, requiring us to determine how to assign label rows to sentences as 1 or 0. In this subsection, we describe the strategies we employed to process the data and bring it to the format illustrated in Figure Table 4.2.

1. **wiki**: In this case, all the values in the label-columns were either 1 or 0, and labels were not mutually exclusive. One sentence could be present in multiple categories, so we set the label rows to 1 if the corresponding label columns contained 1.

2. **measuring_hs**: Interestingly, our research objective differed from the original intent of the dataset creators. We chose to discard information about the "strength" of the labels and reduced them to binary values. Labels in this dataset were not mutually exclusive.

3. **ToLD**: Each category in this dataset had 3 annotators, and each annotator could vote "yes" or "no" (1 or 0) for each category. We considered a sentence belonging to a category if at least 1 annotator voted for it. Labels in this dataset were not mutually exclusive.

4. **measuring_hs**: The original dataset had the column labels "hate_speech," "offensive_language," and "neither." The annotators had to agree upon which category the sentence belongs to. Each vote increased the score of the category, and we assigned the category with the highest score as the label for the sentence. There were no two same highest scores in any row. Labels in this dataset were mutually exclusive.

5. **GermEval**: In this dataset, all sentences had a double gradation. First, they were marked if they were offensive or not, and if marked offensive, they could be categorized into 4 subcategories: "insult," "abuse," "profanity," and "other." The categories "insult," "abuse," and "profanity" became the labels, and labels in this case were mutually exclusive. The raw structure of the dataset is shown in Figure 4.1.

	text	offence	category
1091	Hallo Freunde,bin gerade aus einer poln. Gemei...	OTHER	OTHER
1092	Merkel – Vollbeschäftigung bis 2025, "das ist ...	OFFENSE	INSULT
1093	@user @user @user @politics Hätte Vorschläge, ...	OTHER	OTHER
1094	@user @user @user @user @user Er sagte weiterh...	OTHER	OTHER
1095	@user Die Türken sollen ihr Hass auf alles Nic...	OFFENSE	ABUSE

Figure 4.1: The structure of the GermEval dataset before processing.

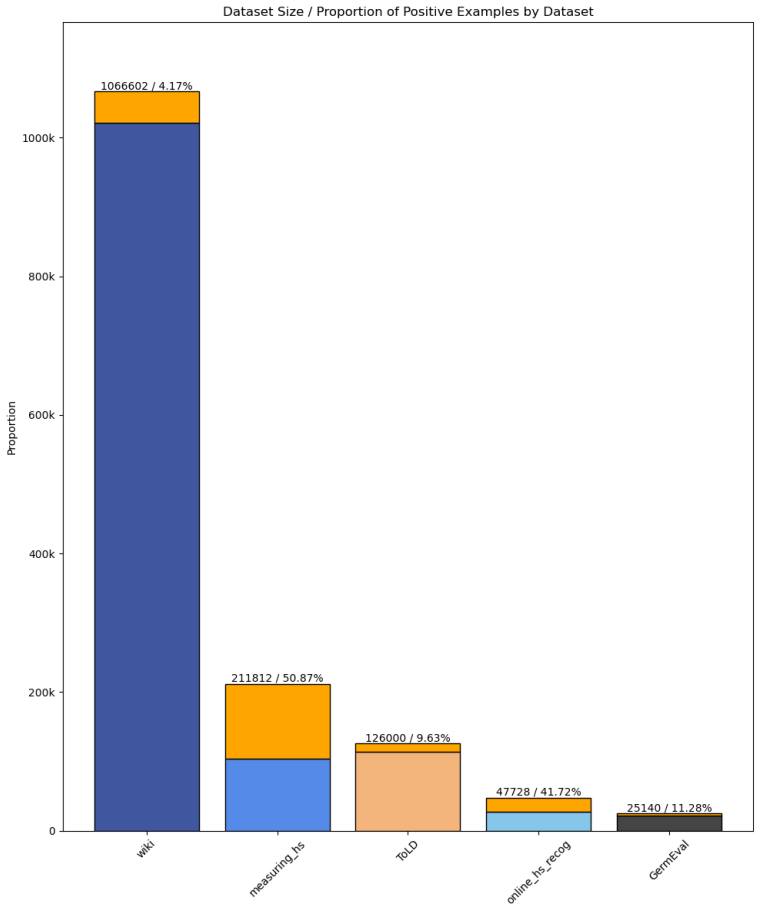

Figure 4.2: The sizes of the datasets and the proportion of sentences positively labeled as hate speech.

5 Settings and Experiments

In this section, we present a comprehensive account of the experiments conducted using the TARS method. Our primary objectives revolved around establishing the method's efficacy when applied to hate speech data, with a specific focus on validating the following properties:

1. **Few- and Zero-Shot Performance:** One of our key objectives was to assess how well the TARS method performed in both few-shot and zero-shot scenarios. In the few-shot setting, we aimed to evaluate the model's performance when exposed to a small amount of labeled data, while in the zero-shot setting, we examined how the model adapted to completely new, previously unseen labels.

2. **Knowledge Accumulation:** Additionally, we explored the TARS method's ability to accumulate knowledge effectively. By exposing the model to increasing amounts of labeled data over time, we sought to determine whether its performance improved with the training progression.

Throughout this section, we will also provide detailed descriptions of the experiment settings, including model configurations, and evaluation metrics used to assess the method's performance on hate speech detection tasks.

5.1 Settings

In this section, we describe the crucial experiment settings, including the coice of the models, hyperparameters, and evaluation metrics used to assess the performance. These settings play a vital role in evaluating the model's effectiveness across various languages and data scenarios.

5.1.1 Models

For our experimental investigations, we employed two Cross-lingual Language Models (CLMs):

1. **Pretrained XLM-RoBERTa-base**: We selected the XLM-RoBERTa-base model due to its prominence as one of the top-performing language models. Its key advantage lies in its ability to understand multiple languages effectively. Given that hate speech and offensive language are prevalent universal phenomena in the online environment, we were particularly interested in leveraging a model with multilingual capabilities.

2. **Pretrained Glot500-base**: Additionally, we utilized the Glot500-base model for our experiments. The rationale behind using this model was its extended multilingual support, catering to an even larger number of languages. Our curiosity lay in discerning potential differences in crosslingual performance between the two models, as hate speech and offensive language transcend linguistic boundaries.

```
batch_size = 6
steps = 2000

training_args = TrainingArguments(
    output_dir= '../checkpoints-xlmr-wiki',
    evaluation_strategy='steps',
    eval_steps=steps,
    learning_rate=2e-5,
    per_device_train_batch_size=batch_size,
    per_device_eval_batch_size=batch_size,
    num_train_epochs=5,
    weight_decay=0.01,
    save_strategy='steps',
    save_steps=steps
    )
```

Figure 5.1: Trainer's arguments.

5.1.2 Hyperparameters

The hyperparameters we used during our experiments were grouped together under the variable "training_args", as illustrated in Figure 5.1.

Due to the varying availability of GPUs during different stages of experimentation, we encountered the necessity to adjust the batch size accordingly. In the best-case scenario, a batch size of 32 was employed, while in the worst-case scenario, it was reduced to 1. Furthermore, constraints on storage space also impacted the selection of the "steps" variable, which represents the total number of steps involved in the training process. Depending on the available storage capacity and the required number of steps, this parameter ranged from 50,000 to 1,000.

For the train-test split, we allocated 80% of the data for training purposes, reserving the remaining 20% for evaluation. This partitioning facilitated the assessment of model performance on unseen data, ensuring a robust evaluation of the TARS method in the context of hate speech detection.

5.1.3 Metrics

Given the highly unbalanced nature of our binary data, we chose to employ F1 scores as the primary evaluation metric for our experiments. The F1 score provides a balanced assessment of precision and recall, making it well-suited for imbalanced datasets. By utilizing this metric, we aimed to accurately gauge the performance of the TARS method in hate speech detection, considering both the ability to correctly identify positive instances (hate speech) and the capability to avoid false positives.

5.2 Experiments

In this section, we provide the details of our experiments, which include different training scenarios and evaluations of the TARS method's performance.

5.2.1 Consecutive Training: Round 1, XLM-RoBERTa

For this experiment, we conducted fine-tuning on the `XLM-RoBERTa-base` model using all five datasets, adding them sequentially, and evaluating the model's performance on each dataset after adding it. The datasets were incorporated into the model in a specific descending order to test the impact of dataset size on the results, following the observation from the original TARS paper that starting training with the largest dataset yielded improved outcomes.

The sequence of dataset incorporation for each step i was as follows:

1. wiki
2. measuring_hs
3. ToLD
4. online_hs_recog
5. GermEval

The exact sequence at step i is as follows:

1. Load the model trained in step $i - 1$.
2. Train the model $i - 1$ using dataset i.
3. Evaluate the model trained in step $i - 1$ on all 5 datasets.

We conducted five consecutive steps, with i corresponding to the dataset ID from 1 to 5. The initial model at step 0 was `XLM-RoBERTa-base`. The final model resulting from this experiment was named "`xlm-r - wiki + measuring_hs + ToLD + online_hs_recog + GermEval`."

5.2.2 Consecutive Training: Round 1, XLM-RoBERTa with Frozen Layers

In this experiment, inspired by a previous study[1], we aimed to investigate the impact of freezing certain layers on the performance of XLM-RoBERTa-base, particularly on the smallest dataset, GermEval. For this purpose, we conducted another round of training with the xlm-r model, but this time we froze specific layers, namely layers 0, 2, 4, 8, 10, 12, and the embedding layer.

The remaining settings for this experiment were kept identical to those of Experiment 1. We performed consecutive training, fine-tuning the model on each dataset, and evaluating its performance after incorporating each dataset.

5.2.3 Consecutive Training: Round 1, Glot500

In this experiment, we replicated the settings of Experiment 1 while using a different model, `Glot500-base`, for Step 0. The goal was to assess its performance in the same consecutive training setup. The settings and methodology were kept exactly the same as in Experiment 1, but this time we used Glot500-base as initial model.

5.2.4 Consecutive Training: Round 2, XLM-RoBERTa

In this experiment we conducted Round 2 of consecutive training with `XLM-RoBERTa-base`, following the exact same procedure as described in Experiment 1. The only difference was that instead of initializing with the base model, we used the model `xlm-r - wiki + measuring_hs + ToLD + online_hs_recog + GermEval` obtained from the end of Round 1 as initial model.

The process and settings remained completely identical to Experiment 1, and the main objective was to replicate the approach taken by the authors of the TARS paper, who conducted 2 rounds of consecutive training on their 5 datasets, as mentioned in Figure 2 in the original TARS paper [1].

[1] https://raphaelb.org/posts/freezing-bert/

5.2.5 Simultaneous Training: XLM-RoBERTa

After witnessing promising results and consistently improving average F1-scores during consecutive training, we were intrigued to explore the impact of feeding the datasets simultaneously rather than consecutively. To investigate this, we merged all processed datasets and performed a single run of training `xlm-r` on the combined data.

Our motivation for this experiment was to assess the potential performance gains or differences resulting from simultaneous training, compared to the consecutive training method employed in the previous experiments.

The evaluation process remained unchanged: we tested the model on the test portions of all five datasets independently.

5.2.6 Few-Shot (k-Shot) Learning with XLM-RoBERTa

In our final experiment, we aimed to test the few-shot learning capacity of the TARS-method by varying the k value (1, 2, 4, 8, 10, and 100). Here, k represents the sample size, indicating the number of instances used for fine-tuning. We used the model obtained in the first step of the first run, `XLM-RoBERTa-base` trained on the `wiki` dataset.

For each k value, we loaded the model, performed fine-tuning on a slice of the `measuring_hs` dataset with a size equal to k, saved the updated model and conducted evaluation on all five datasets.

6 Result Analysis

In this chapter, we present the F1 scores obtained from the experiments conducted in Chapter 5. We will discuss the performance of the TARS method in hate speech detection using the datasets wiki, measuring_hs, ToLD, online_hs_recog, and GermEval. The evaluation metrics used in the experiments will be detailed, and we will analyze the tendencies and patterns discovered during the evaluation.

6.0.1 Consecutive Training: Round 1, XLM-RoBERTa

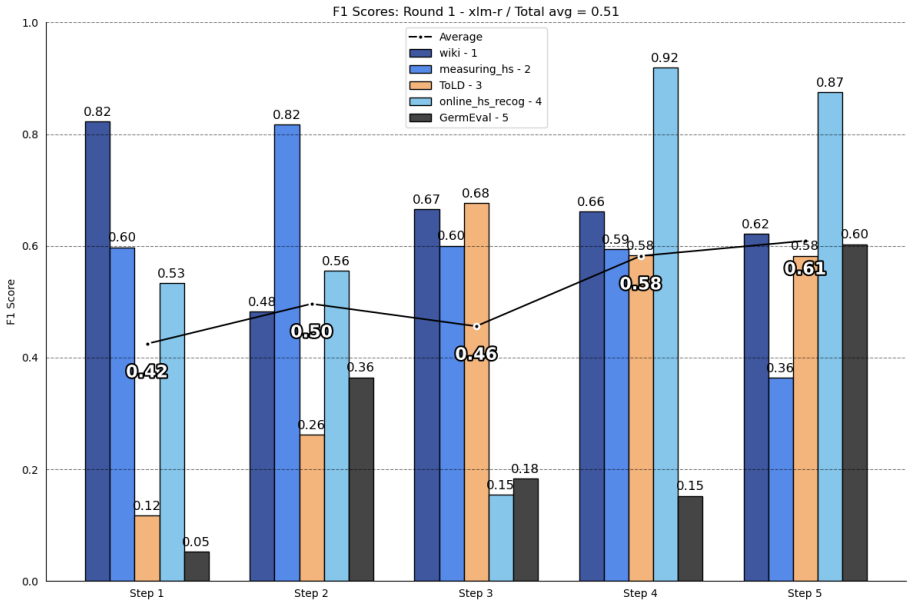

Figure 6.1: F1 Scores: Round 1 - XLM-R.

Figure 6.1 illustrates the results of consecutive training with the `XLM-RoBERTa` model. Notably, the step-average F1 score demonstrates an increasing trend, indicating that the TARS method can effectively retain knowledge when trained on multiple datasets sequentially. The overall average F1 score for all steps is 0.51.

Upon closer examination, we observe that the increase in F1 scores is not linear, as individual dataset performance fluctuates significantly with each step. Specifically, datasets `wiki` (Step 1), `measuring_hs` (Step 2), and `online_hs_recog` (Step 4), all in English, show performance improvements during the initial steps. However, a notable decline in the average F1 score occurs at Step 3 when the model is trained on the Portuguese dataset, `ToLD`. Similarly, after training on the German dataset, `GermEval`, at Step 5, there is a decrease in the average F1 score for the English datasets at Step 4.

The dynamics shown in Figure 6.2 suggest that the difference in languages between datasets may contribute to performance variations during consecutive training. Further analysis is required to explore the impact of linguistic diversity on the TARS method's performance during training rounds.

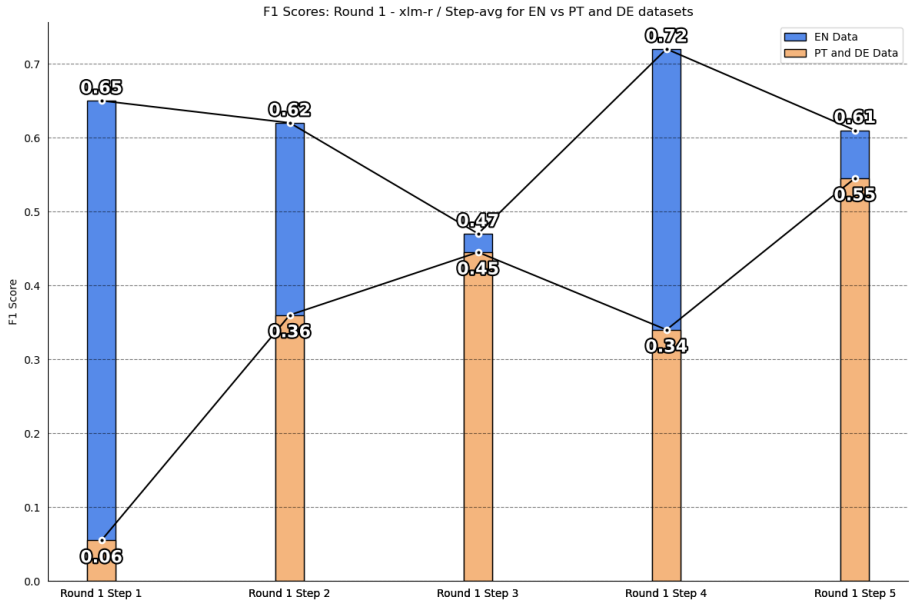

Figure 6.2: Counteracting Dynamics of EN Data vs PT and DE Data.

6.0.2 Consecutive Training: Round 1, XLM-RoBERTa with Frozen Layers

Figure 6.3 depicts the results of consecutive training with the `XLM-RoBERTa` model while freezing certain layers. Surprisingly, contrary to our expectations, freezing layers did not lead to performance improvements. Instead, the model with frozen layers yielded lower F1 scores for almost every step, with an average F1 score of 0.49, which was 0.2 worse than the model without freezing.

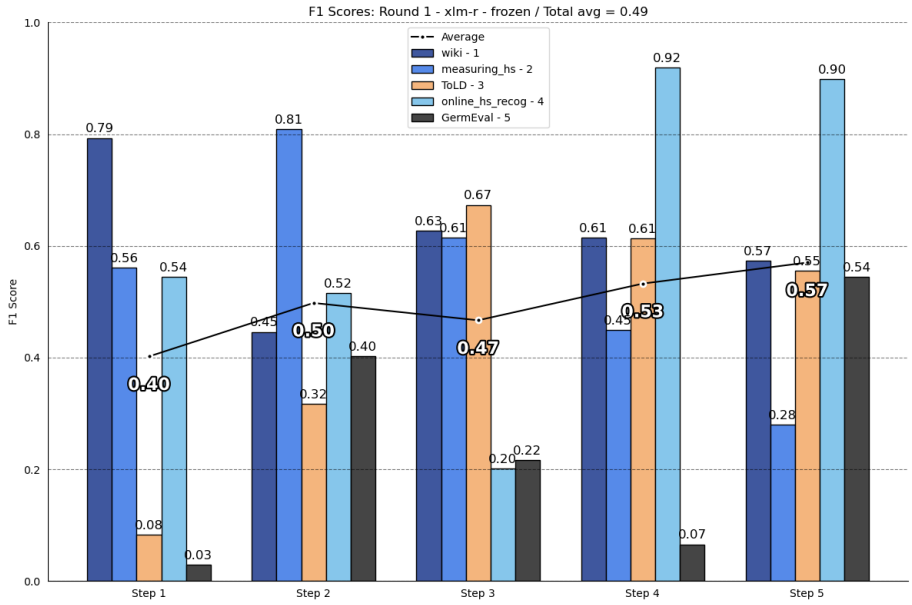

Figure 6.3: F1 Scores: Round 1 - XLM-R with Frozen Layers.

Similar to the previous experiment, we observe the same dynamic of fluctuations, but with a lower magnitude. The disturbance in performance occurs at Step 3 when the model is trained on the Portuguese dataset, `ToLD`, and there is a subsequent fall in F1 scores for the English datasets at Step 5, after incorporating the German dataset, `GermEval`. These findings suggest that freezing layers may not be beneficial for the TARS method during consecutive training, and further investigation is needed to understand the reasons behind

this observation.

6.0.3 Consecutive Training: Round 1, Glot500

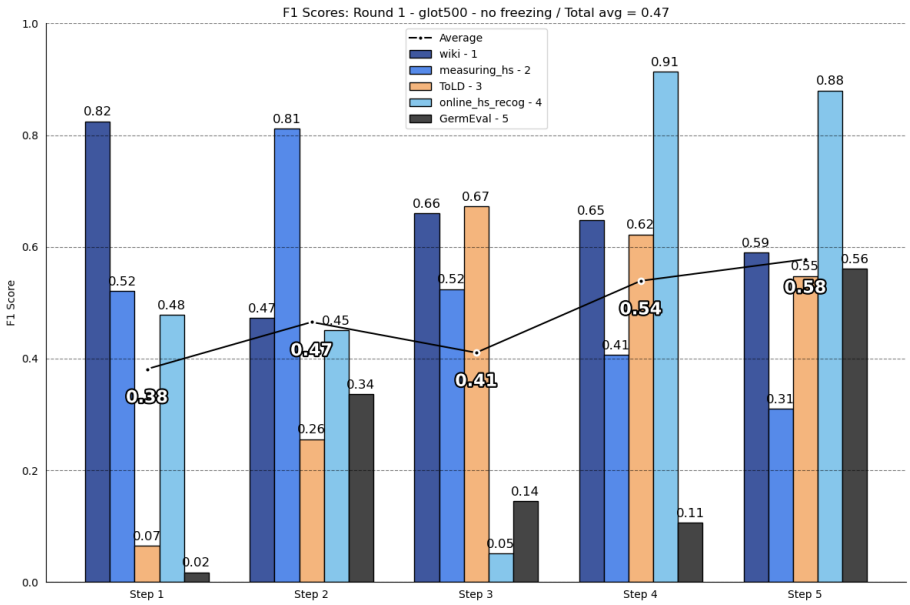

Figure 6.4: F1 Scores: Round 1 - Glot500.

Figure 6.5 presents the results of consecutive training with the `Glot500-base` model. The performance dynamic of the Glot500 model closely mirrors that of the XLM-RoBERTa model. However, one notable difference is the consistent drop in performance at every step.

6.0.4 Consecutive Training: Round 2, XLM-RoBERTa

In the second round of consecutive training with `XLM-RoBERTa`, we observed a slight shift in the F1 score dynamics compared to the first round, as shown in Figure 6.5.

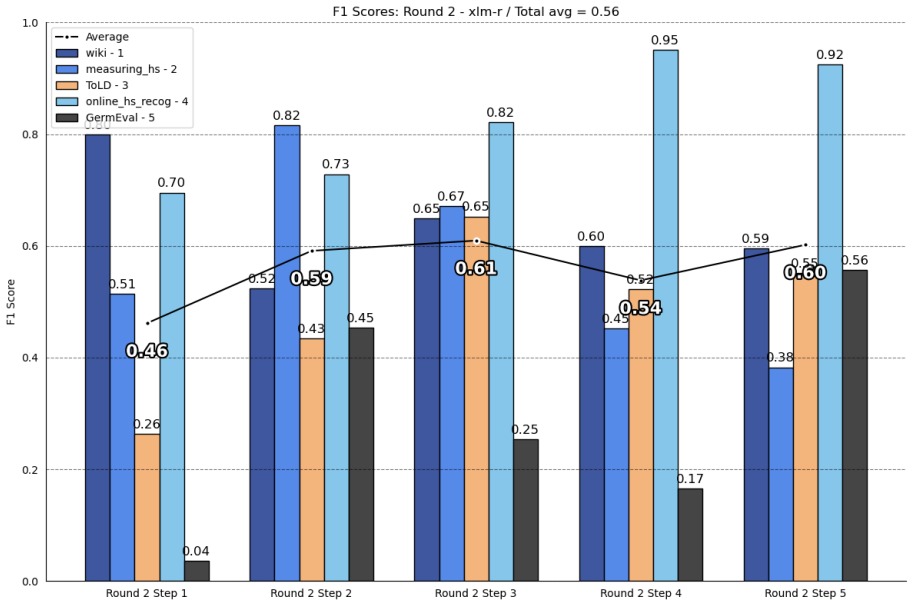

Figure 6.5: F1 Scores: Round 2 - XLM-R.

One prominent observation is the significant fluctuations in the F1 score for the GermEval dataset. After training the model on the wiki dataset for the second time, the F1

score for GermEval experienced a notable drop.

Another noteworthy point is that, in the second round, the spikes and drops in the step-average F1 scores do not directly correspond to the languages of the datasets. Further analysis is necessary to understand the reasons behind these fluctuations and variations in the second round of training.

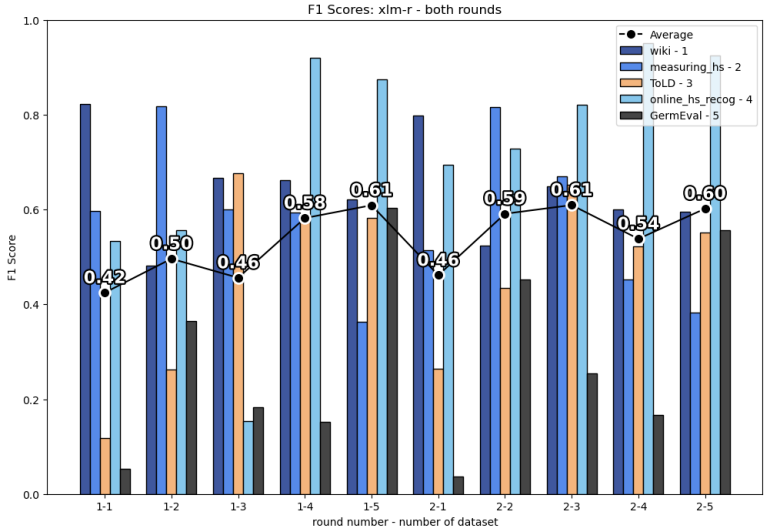

Figure 6.6: F1 Scores: Both Rounds - XLM-R.

The F1 score dynamics for both rounds can be seen in Figure 6.6.

6.0.5 Simultaneous Training: XLM-RoBERTa

After conducting a series of experiments with consecutive training, we were intrigued to explore the model's performance when the entire data was fed at once. The results shown in Figure 6.7 were quite surprising: The average F1 score of 0.86 and the lowest individual F1 score of 0.77 outmatched by a large margin everything we achieved during the consecutive training experiments. Additionally, we did not observe any significant discrepancies between the datasets, unlike what we observed during consecutive training.

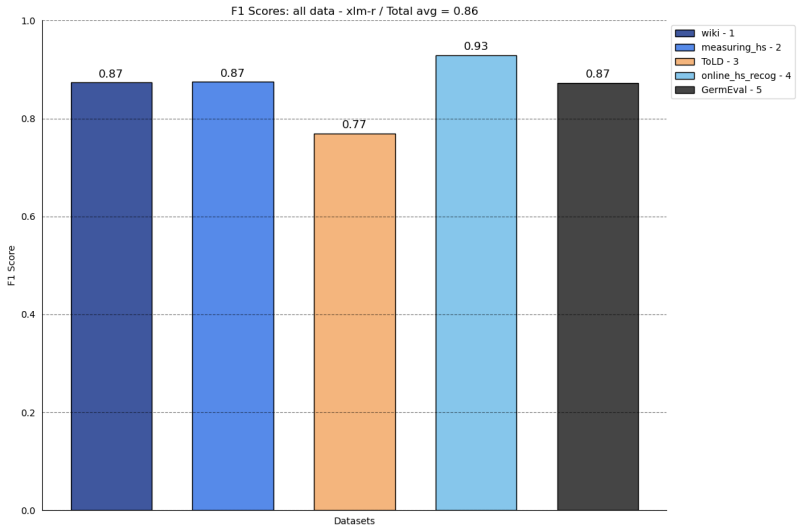

Figure 6.7: F1 Scores: XLM-R Simultaneous Training.

Notably, this model's performance surpassed the results reported by the research teams that produced the ToLD and GermEval datasets. For both datasets, the best F1 scores

achieved by models specifically built for them were 0.76 and 0.77, respectively. The outstanding performance of the model under simultaneous training suggests the effectiveness of this approach for hate speech detection tasks, outperforming specialized models built for individual datasets. Further investigation is warranted to explore the reasons behind this remarkable performance gain.

6.0.6 Few-Shot (k-Shot) Learning with XLM-RoBERTa

The original TARS paper explored k-shot learning with various k values (0, 1, 2, 4, 8, 10, 100) in the context of different tasks, including sentiment analysis. They reported a slight overall increase in performance with the increasing k-value, as demonstrated in Table 4 of the original paper. Our primary motivation was to test the model under similar conditions using our data and observe if we would observe a similar performance trend and improvement.

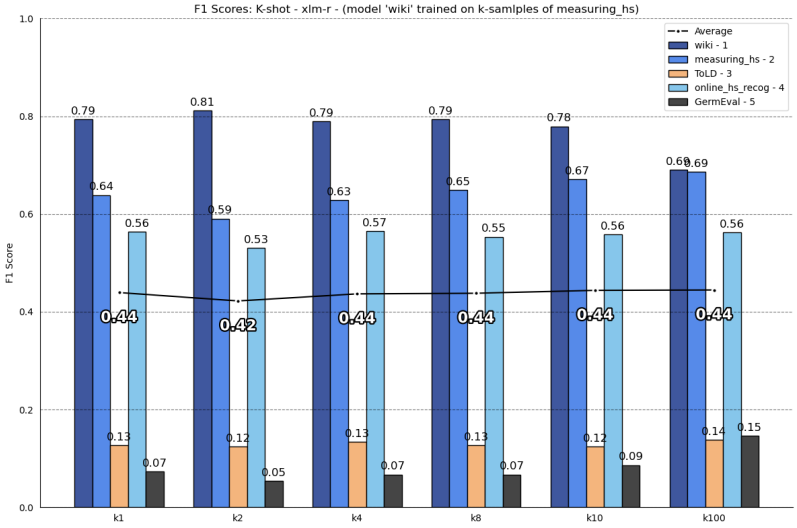

Figure 6.8: F1 scores: Impact of k-value on the model performance.

We excluded k=0 from our experiments, as it was already extensively researched in our previous experiments. Surprisingly, contrary to our expectations, we did not observe any significant improvements in the step-average F1 score with an increasing value of k. However, we did notice a consistent tendency similar to the consecutive training experiments: As the F1 scores of the Portuguese and German datasets grew with the increasing k-value, those of the English datasets decreased. This tendency became more pronounced as the k-value increased, as can be seen in Figure 6.8.

7 Conclusion

In this chapter, we present our conclusions based on the extensive experiments conducted in the previous sections. The experiments encompassed various training scenarios and evaluations, aiming to explore the performance and effectiveness of the TARS method in hate speech detection. Here, we summarize the key findings and insights obtained from our research.

7.1 Knowledge Retention and Zero-Shot Capabilities

In Sections 5.2.1 to 5.2.3, we conducted consecutive training with different models and observed remarkable zero-shot capabilities of the TARS method. After just the first step of training, where only one English dataset was used, the `XLM-RoBERTa-base` model achieved decent F1 scores of **0.60** and **0.53** on two other English datasets. Throughout both rounds of consecutive training, we observed the model's ability to retain previously learned knowledge. The step-average F1 score increased from **0.42** in the first step of the first round to **0.60** at the last step of the second round. Although the increase was non-linear with significant spikes and declines, the overall trend of the step-average F1 score to increase was evident, with a round-average F1 score of **0.51** in round 1 and **0.56** in round 2. Comparing the three models in round 1, we observed that `XLM-RoBERTa-base` without freezing performed the best, yielding a total average F1 score of **0.51**, while `Glot500-base` performed the worst.

7.2 Impact of different languages

In all experiments involving consecutive training, we observed that the language of the dataset had a noticeable effect on the model's performance. There was a clear tendency where the Portuguese and German datasets had a negative impact on the progress achieved on the English datasets, and vice versa. This observation is novel, as the original TARS research only utilized data in English. The interaction of different languages within such a model requires further research to better understand its implications and potential improvements.

7.3 Simultaneous training

The most outstanding result was achieved when all datasets were simultaneously fed into the model as a single chunk, yielding a total average F1 score of **0.86**. Notably, this approach not only produced the best result among all the experiments but also showed no obvious counteraction between the English, Portuguese, and German datasets.

7.4 Few-shot (k-shot) learning

In the context of few-shot (k-shot) learning, our experiments revealed that on our data, the choice of small k-values did not appear to exert a significant influence on the overall performance. Despite varying the k-value from 1 to 100, we did not observe substantial improvements in the step-average F1 scores. The model's ability to leverage limited instances for fine-tuning did not lead to notable performance gains in our experimental setup. These

findings suggest that the few-shot learning capacity of the TARS method may have limitations when applied to hate speech detection using our specific datasets. However, further investigation may be warranted to explore the potential impact of different k-values on other datasets or tasks.

7.5 summary

In summary, our experimental findings affirm the reported properties of the TARS method by its authors, showcasing its impressive zero-shot capabilities and knowledge retention. These results clearly demonstrate the potential of the TARS method as a robust foundation for collaborative efforts within the community, facilitating the seamless integration of new data into the model. The adaptability and performance showcased in our experiments underscore the effectiveness of the TARS method for hate speech detection. Furthermore, the flexibility offered by the TARS method opens up the possibility of widespread community cooperation, where different teams can readily contribute by providing new data or further continuing the model training. This collaborative approach can enhance the model's generalization and adaptability across different datasets and languages, making it a valuable asset for addressing hate speech and offensive language in various online contexts.

8 Future work

In the future, there are several intriguing possibilities to further investigate the potential of the TARS approach:

1. Testing this method on a broader range of languages to understand how different languages interact and whether there are instances where various languages complement each other. Additionally, exploring the impact of multilingual labels could provide valuable insights.

2. Experimenting with consecutive training on monolingual data in multiple languages to assess its effectiveness compared to the multilingual approach.

3. Training `XLM-RoBERTa` with frozen layers for a larger number of epochs to determine if extended training enhances performance.

4. Conducting additional rounds of consecutive training with all three models to observe if repeated iterations yield improved long-term results.

5. Exploring diverse data adding configurations, such as combining consecutive training on one group of datasets with simultaneous training on another group, to identify the most effective approaches for cooperation and knowledge sharing among different studies.

These future endeavors will help advance our understanding of the TARS method's capabilities and shed light on its potential applications in hate speech detection across various languages and datasets.

Bibliography

[1] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online, August 2021. Association for Computational Linguistics.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.

[3] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[5] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51:1 – 30, 2018.

[6] Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. Task-aware representation of sentences for generic text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3202–3213, 2020.

[7] Chris J. Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *CoRR*, abs/2009.10277, 2020.

[8] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019.

[9] João Augusto Leite, Diego F. Silva, Kalina Bontcheva, and Carolina Scarton. Toxic language detection in social media for brazilian portuguese: New dataset and multilingual analysis. *CoRR*, abs/2010.04543, 2020.

[10] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[11] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[13] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537, 2019.

[14] Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. Overview of the germeval 2018 shared task on the identification of offensive language. 09 2018.

[15] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914, 2016.

[16] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *CoRR*, abs/1903.08983, 2019.

List of Figures

4.1 The structure of the GermEval dataset before processing. 20
4.2 The sizes of the datasets and the proportion of sentences positively labeled as hate speech. 20

5.1 Trainer's arguments. 22

6.1 F1 Scores: Round 1 - XLM-R. 25
6.2 Counteracting Dynamics of EN Data vs PT and DE Data. 26
6.3 F1 Scores: Round 1 - XLM-R with Frozen Layers. 26
6.4 F1 Scores: Round 1 - Glot500. 27
6.5 F1 Scores: Round 2 - XLM-R. 27
6.6 F1 Scores: Both Rounds - XLM-R. 28
6.7 F1 Scores: XLM-R Simultaneous Training. 28
6.8 F1 scores: Impact of k-value on the model performance. 29

List of Tables

4.1 Short description of the datasets used in this research. 17
4.2 The end-format of the data we used in the experiments. 18
4.3 An example of a raw sentence taken from the wiki-dataset. 18
4.4 The formatted version of the sentence from Table 4.3. 18
4.5 An example of a formatted sentence from a non-harmful category from the GermEval dataset. 19