Višejezično otkrivanje govora mržnje i uvredljivog govora

Iveta, Mateja

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS No. 686

MULTILINGUAL OFFENSIVE AND HATE SPEECH DETECTION

Mateja Iveta

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

MASTER THESIS No. 686

MULTILINGUAL OFFENSIVE AND HATE SPEECH DETECTION

Mateja Iveta

Zagreb, June 2024

UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

Zagreb, 04 March 2024

MASTER THESIS ASSIGNMENT No. 686

Student:	**Mateja Iveta (0036513744)**
Study:	Computing
Profile:	Computer Science
Mentor:	prof. Jan Šnajder

Title: **Multilingual offensive and hate speech detection**

Description:

Content moderation plays a crucial role in today's social media platforms, helping to combat hate speech, harassment, and inappropriate content. However, existing research in this area primarily focuses on a limited set of high-resource languages or on English and one other lower-resource language. This limited scope neglects the diversity of languages used online, highlighting the need for broader research encompassing various linguistic backgrounds. The objective of this thesis is to advance hate speech detection by leveraging data from multiple languages, with the goal of developing a multilingual model for detecting profanity, offensive language, and hate speech. Utilize publicly available multilingual data to train a robust multilingual offensive and hate speech detection model. Investigate alternative architectural approaches and model combinations to enhance model performance, particularly in handling domain-specific hate speech. All references must be cited, and all source code, documentation, executables, and datasets must be included with the thesis.

Submission date: 28 June 2024

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 4. ožujka 2024.

DIPLOMSKI ZADATAK br. 686

Pristupnica:	**Mateja Iveta (0036513744)**
Studij:	Računarstvo
Profil:	Računarska znanost
Mentor:	prof. dr. sc. Jan Šnajder

Zadatak: **Višejezično otkrivanje govora mržnje i uvredljivog govora**

Opis zadatka:

Moderiranje sadržaja igra ključnu ulogu na današnjim platformama društvenih medija pružajući pomoć u borbi protiv govora mržnje, uznemiravanja i neprimjerenog sadržaja. Međutim, postojeća istraživanja u ovom području usredotočena su prvenstveno na ograničeni skup jezika sa značajnim resursima ili na engleskom uz jedan dodatan jezik s malo resursa. Ovaj ograničeni opseg zanemaruje raznolikost jezika koji se koriste na internetu, naglašavajući potrebu za širim istraživanjem koje obuhvaća različita jezična podrijetla. Cilj ovog rada jest unaprijediti otkrivanje govora mržnje korištenjem podataka iz više jezika, s ciljem razvoja višejezičnog modela za otkrivanje vulgarnosti, uvredljivog jezika i govora mržnje. Iskoristiti javno dostupne višejezične podatke za učenje robusnog višejezičnog modela za otkrivanje uvredljivosti i govora mržnje. Istražiti alternativne arhitekture i kombinacije modela kako bi se poboljšala izvedba, posebno u rješavanju govora mržnje specifičnog za određenu domenu. Radu priložiti izvorni i izvršni kod radnog okvira, skupove podataka i programsku dokumentaciju te citirati korištenu literaturu.

Rok za predaju rada: 28. lipnja 2024.

Contents

1. Introduction

The issue of hate speech is an everyday issue that makes an impact on both individual and global levels. While its definition has many variations, it is, in essence, an offensive discourse about an individual or a group based on some inherent characteristics.[1] Generally, it is rooted in stereotyping and stigmatisation based on identity factors such as race, religion, and skin colour, gender or others. That kind of expression can spark severely negative and sometimes even dangerous consequences.

With social media becoming something almost everyone participates in, the anonymity and reach it provides have only fuelled opportunities for animosity. It may solidify or intensify opinions about certain groups with exposure to many like-minded people, enticing them to join the discourse. In extreme cases, it can be used as a platform for a public call for violence. This is not to exclude traditional media and in-person means of spreading hate speech. However, the threshold for entry into social media spaces is much lower.

There is an increasing need for efficient content moderation strategies to, ideally, stop the rapid spread of hate and misinformation. It is important to accept the shortcomings of human moderation – speed and uniformity. With larger platforms like Twitter, Reddit, and Meta, the sheer amount of content created every day cannot be managed manually. Another thing to consider is that not all people view the same types of comments as equally offensive. That all led to the focus shifting towards automated detection – from the more basic blacklisting methods to using large language models for the task.

Natural Language Processing (NLP) is a subfield of computer science and artificial intelligence that enables computers to better understand human language by using machine learning methods. Large Language Models (LLMs) are foundation models trained on large amounts of data, making them capable of dealing with natural language in a wide array of tasks.

This thesis aims to use NLP and LLMs to detect hate and toxic speech text. More specifically, look into detection in multiple languages. The datasets used were English,

[1] `https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech`

Croatian, Indonesian, French, and Portuguese. Chapter 2 provides an overview of the work done on this topic in competitions and research and a basic overview of the BERT model architecture. Then, in Chapter 3, we provide a more detailed description of the datasets and an analysis. Chapter 4 introduces all the models and methods used, explains how the models and datasets were adapted, and describes the evaluation metrics used for model comparisons. In Chapter 5, we show the experiments and their respective results, while the last chapter, Chapter 6, discusses the achieved results and possible future improvements for the project.

2. Background and Related Work

The research area of hate speech, toxic speech, and profanity detection has gained more traction in the recent years more than ever before. This type of language has always been present in everyday life and traditional media but has accelerated rapidly with the availability of information about politically and socially sensitive and divisive events on the Internet. It can manifest in the form of cyberbullying and harassment on forums, and can even translate into large-scale violence. The anonymity, albeit limited, provided by social media can also be considered one of the enabling factors for the general public's boldness in expressing every opinion no matter how harmful it may be. It is also important to acknowledge the benign uses of profane speech as a form of expression.

With these components taken into consideration, the number of papers on this topic

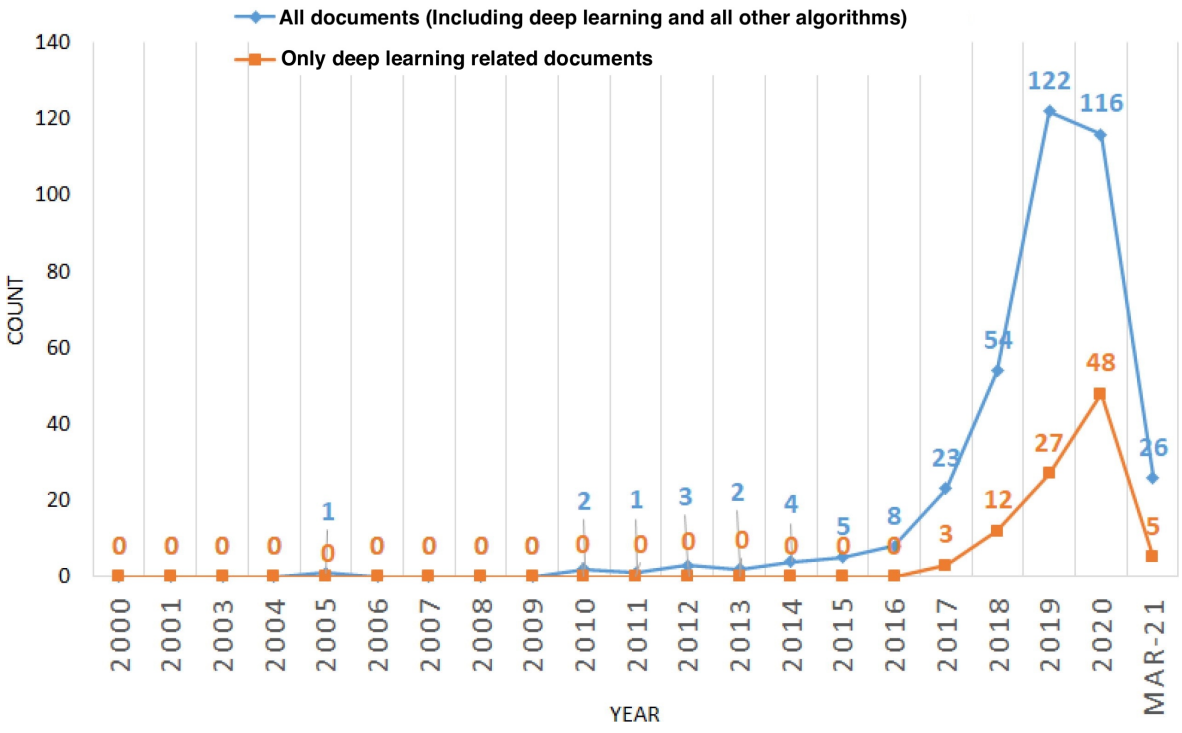

Figure 2.1: Number of publications per year from 2000–2021 related to automatic hate speech detection in NLP. Figure taken from (Jahan and Oussalah, 2023).

has been increasing, peaking in 2019 as Figure 2.1 illustrates. The majority of research has focused on high-resource languages such as English, Chinese, and Spanish. This comes as no surprise due to several factors. Firstly, the amount of data available for these languages is simply larger. They have a significant number of speakers, native or otherwise, so more content is being produced in these languages. English, for example, along with many native speakers, has unofficially become the language of the internet and the research community. Consequently, the foundation for experimentation is more available: with most of the popular models having been trained in these languages, it is easier to continue upgrading instead of building from scratch. Lastly, and similarly to the first point, frequent usage leads to a higher impact and usefulness of such solutions.

2.1. Related work

Significant efforts have already been made to create a systemic overview of the topic encompassing the analysis of the available datasets and the popular methods and models used. This has provided insight into the languages and sourcing distributions. Most notably, Jahan and Oussalah (2023) have presented research that shows the commonalities between the approaches. Unsurprisingly, as can be seen in Figure 2.2, looking only at publicly available datasets, they confirmed that the strong majority of research has been done on English data. Another trend, shown in Figure 2.3, was the prevalence of social media as the source for dataset creation. Sources like that, while readily available and expansive, contain noisy data due to the users adapting to dodge the automated detection filters. They state that the transformer architectures have generally outperformed other methods with the rise of Bidirectional Encoder Representations from Transformer (BERT) as defined by Devlin et al. (2018). BERT is a deep-learning model designed for understanding natural language, trained on large text corpora enabling it to grasp the context from both directions. The model, and its variants, are further described in Section 4.2.

A similar study was done a couple of years earlier by Vidgen and Derczynski (2021), which concluded that creating large, varied, and minimally biased datasets for the task of hate speech (and all related subcategories) detection is a difficult feat. They also state that combining the existing data into a larger-scope dataset is unlikely due to the specificity of each dataset, whether that arises from the different languages or, for example, the subtlety levels.

2.1.1. Literature overviews

Chhabra and Vishwakarma (2023) provide a comprehensive literature survey on the

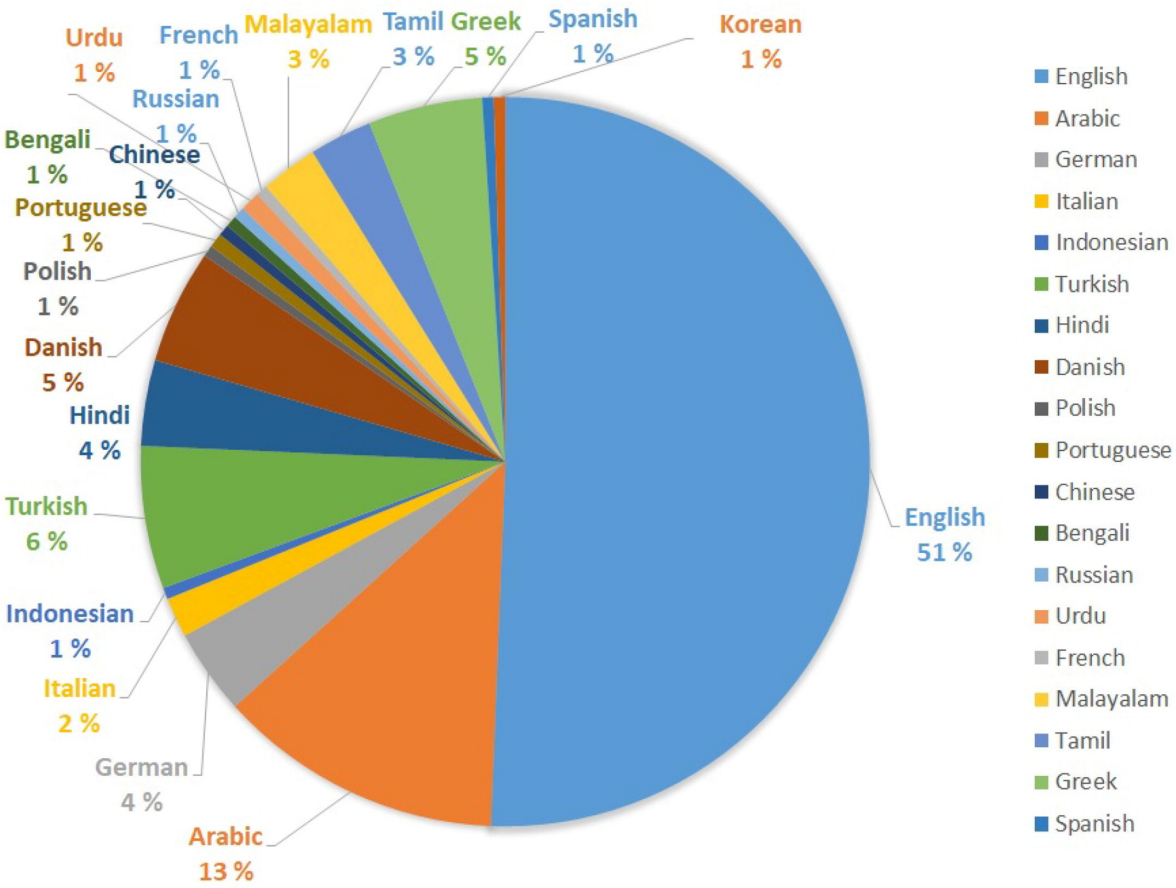

Figure 2.2: Percentage of previous work on hate speech with respect to different languages. Figure taken from (Jahan and Oussalah, 2023).

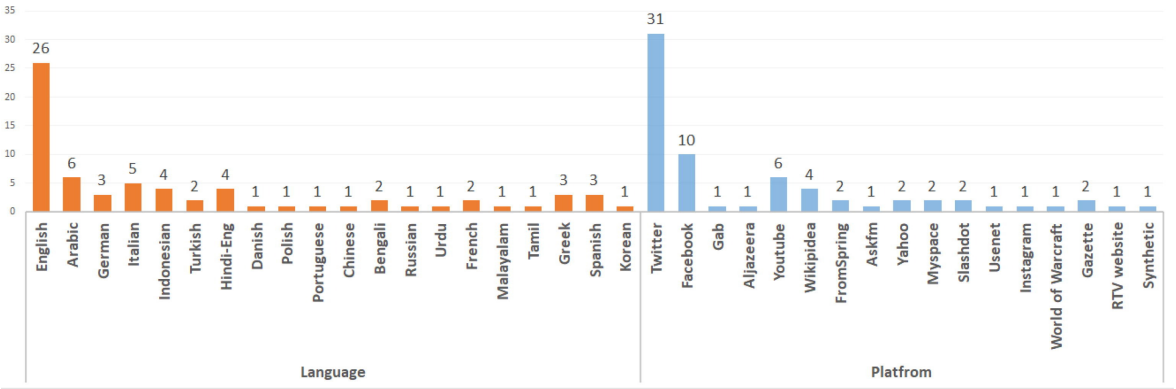

Figure 2.3: Distribution of available hate speech datasets per language and source. Figure taken from (Jahan and Oussalah, 2023).

topic of multilingual automatic hate speech identification. The study encompasses over 120 articles based on keywords such as "hate speech detection". Their work focuses on comparing the methods used in the studies to date and attempts to find and define benchmark datasets and models.

One of the findings is the presence of unlabelled or unreleased non-English datasets and the prevalence of code-switching. Code-switching, or language alternation occurs when a speaker uses two or more languages in a single sentence or a conversation. When it comes to benchmark models, they find that most of them are based on large language models. However, another notable conclusion is the lack of standardised evaluation metrics, which complicates performance comparison and identification of best practices. Röttger et al. (2022), recognising a need for a more unified testing method created `MULTILINGUAL HATE CHECK` (MHC), which is a suite of functional tests for multilingual hate detection models for 10 languages.

A year prior, Alkomah and Ma (2022) also provided an overview of the results from relevant papers showing the different metrics used to benchmark their results. Their study analysed 138 articles and found that several approaches have inconsistent performance within various hate speech categories. Overall, they conclude that 29% of solutions are hybrid – a combination of two or more different machine learning methods. They also find that the granularity and overlap of hate and toxic speech impact the ability to achieve good results. This hierarchy can be seen in Figure 2.4. Directly offensive texts are more easily classified than more broadly hateful texts.

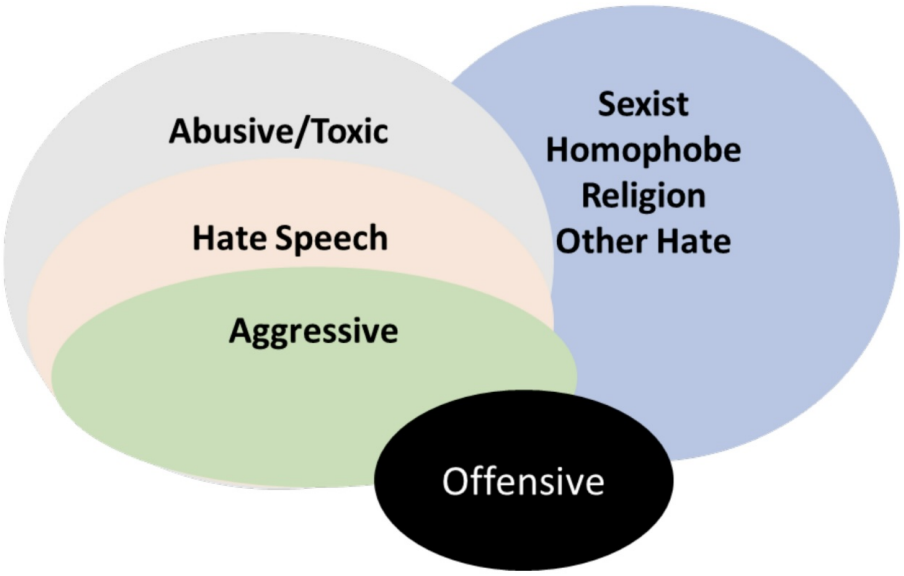

Figure 2.4: Hierarchy of hate speech concepts. Figure taken from (Alkomah and Ma, 2022).

2.1.2. SemEval

The renowned NLP contest SemEval has featured an offensive speech-related back-to-back task in the past five years which can also be interpreted as an indicator of the topic's relevance in the academic community. The first such task as defined by Zampieri et al. (2019) encompassed the identification of offensive language on social media in English. Out of the top 10 teams in the first sub-task – offensive language classification – seven of them used the BERT model. The following year brought an extension to multiple languages (Zampieri et al. (2020)) showcasing the need for a more global approach. Similarly to the previous year, all top 10 teams used BERT, RoBERTa, or XLM-RoBERTa models.

In the last two years, the tasks have focused more on the subtleties, such as in Perez-Almendros et al. (2022), which looked into condescending language, and the finer granulation of online sexism as proposed by Kirk et al. (2023). This required the models to make more nuanced interpretations and to rely on context.

Lastly, Pavlopoulos et al. (2021) created a task that looked into the detection of toxic spans and transformed it into a sequence labelling problem which added a layer of complexity, but they also generated the general toxicity labels. The dataset created for that task is one of the datasets used in this thesis. Because the tasks involve different approaches, sequence labelling versus text classification, we can't compare the results to the ones achieved by the participants.

This iterative refinement and expansion of the task, along with creating more datasets and encouraging research on the topic, is a good sign of its relevance and interest.

2.1.3. Similar studies

The paper by Haidar et al. (2017) tackles the related issue of cyberbullying detection in Arabic, stating that the complex morphological nature of Arabic could partially be the reason for the scarcity of similar NLP research. They utilised shallow learning methods such as Naive Bayes and support vector machines (SVMs). The achieved results were overall good, with an F-score of 92.7%, but due to the large class imbalance, the results for the underrepresented class were lower than the non-cyberbullying class with over 60% of false positives.

Aluru et al. (2020) investigated the comparison of multilingual hate speech detection in nine languages from sixteen sources. They found that BERT-based models perform better in higher-resource cases. Conversely, in low-resource scenarios simpler models performed better, with more effective generalisation capabilities. In some languages,

such as Italian and Portuguese, zero-shot classification proved to be a viable alternative. Moon et al. (2020) also show that the BERT model outperforms the traditional baselines in Korean.

Ranasinghe and Zampieri (2021) employed a similar approach by applying cross-lingual contextual embeddings and transfer learning for seven lower-resource languages. They developed solutions based on popular transformer models – BERT and XLM-RoBERTa. Their results demonstrate that, even with limited training data, there is an option that can perform well, in their case XLM-RoBERTa with transfer learning. Rizwan et al. (2020) concluded that transfer learning is more beneficial for the task than training the embeddings from scratch in Roman Urdu.

Toraman et al. (2022) explore cross-domain transfer learning with English, a high-resource language, and Turkish, a low-resource one. They show that, in such case, transformer-based models outperform conventional bag-of-words models by up to 10%, and that over 90% of the performance capabilities can be retained by using only 20% of the 100,000 examples in their dataset.

2.2. BERT

Bidirectional Encoder Representations from Transformers (BERT) is a language model introduced by Devlin et al. (2018) as a model that significantly outperformed many previous state-of-the-art solutions. It is designed to pre-train deep bidirectional representations – capturing the context of a word by looking at words that come before and after, enabling it to be fine-tuned for a task with just one additional output layer. This section describes the foundational components and the pretraining processes of BERT. Embeddings are trained numerical representations of categorical features. In NLP, it is essentially text-to-vector mapping, which enables the model to process textual data. For BERT, WordPiece tokenization is first applied, breaking down the text into a list of integer token IDs representing a word or a part of the word. Afterwards, those tokens are passed down to the embedding layer. The initial BERT embedding layer consists of three types of embeddings: token embeddings, position embeddings, and segment embeddings.

- **Token embeddings** are the embeddings of each word or subword in the BERT Tokenizer's vocabulary. The approach allows it to directly represent commonly used words, but if a word is unknown, it can still be mapped into a vector as multiple familiar tokens.
- **Position embeddings** represents the position of the tokens in the input sentence. There are 512 different position embeddings due to BERT's input se-

quence limit of 512 tokens.

– **Segment embeddings**, later better known as Token Type Embeddings, are used for the Next Sentence Prediction task – they denote whether the tokens come from sentence `A` or `B` that logically follows sentence `A`.

Figure 2.5 shows the entire BERT embedding layer with the three previously described types of embeddings.

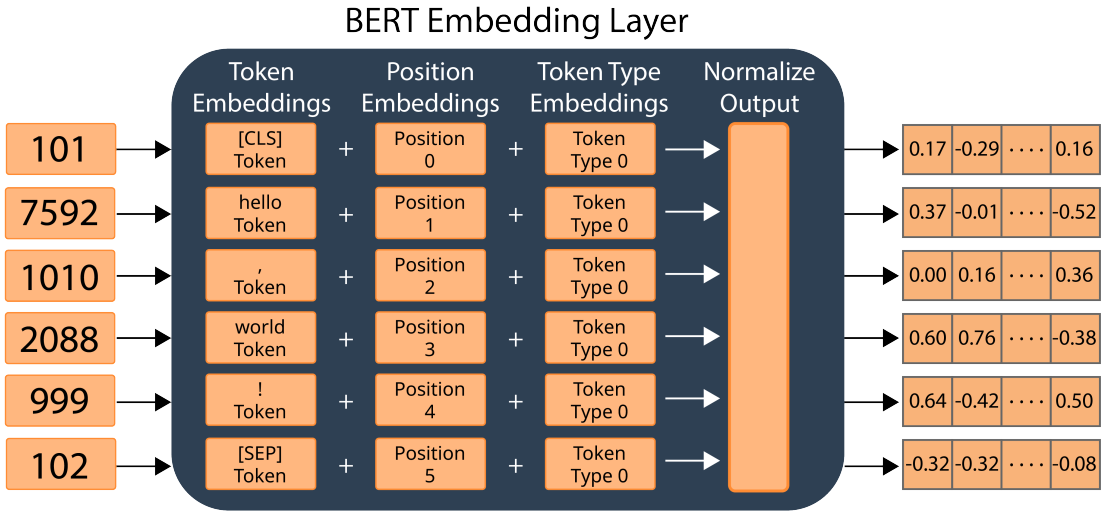

Figure 2.5: BERT Embedding layer. Figure taken from Tinkerd (2023).

2.2.1. Attention Mechanisms

Attentions are a way to assign a weight to inputs based on their importance for the task. For example, if the task is determining if a movie review is positive, words like "`good`" and "`bad`" might be more indicative than "`dog`". It can also be used to form connections between words.

Self-attention is a central mechanism in the transformer architecture of BERT. For each word, the model computes the similarity between that word and all other input words, letting it know what words might be important. The similarities are then turned into attention scores or weights that modify the output values. This method includes information from the entire sentence for every word. Most commonly, it is a scaled dot-product of the inputs.

The multi-head attention process performs the self-attention multiple times in parallel. Each of the separate self-attentions is called a head, which can then focus on a different subspace and capture some different linguistic features. Their outputs are concatenated and linearly transformed to produce the final representation. Figure 2.6

shows the multi-head attention layer and how it is comprised of multiple self-attention heads.

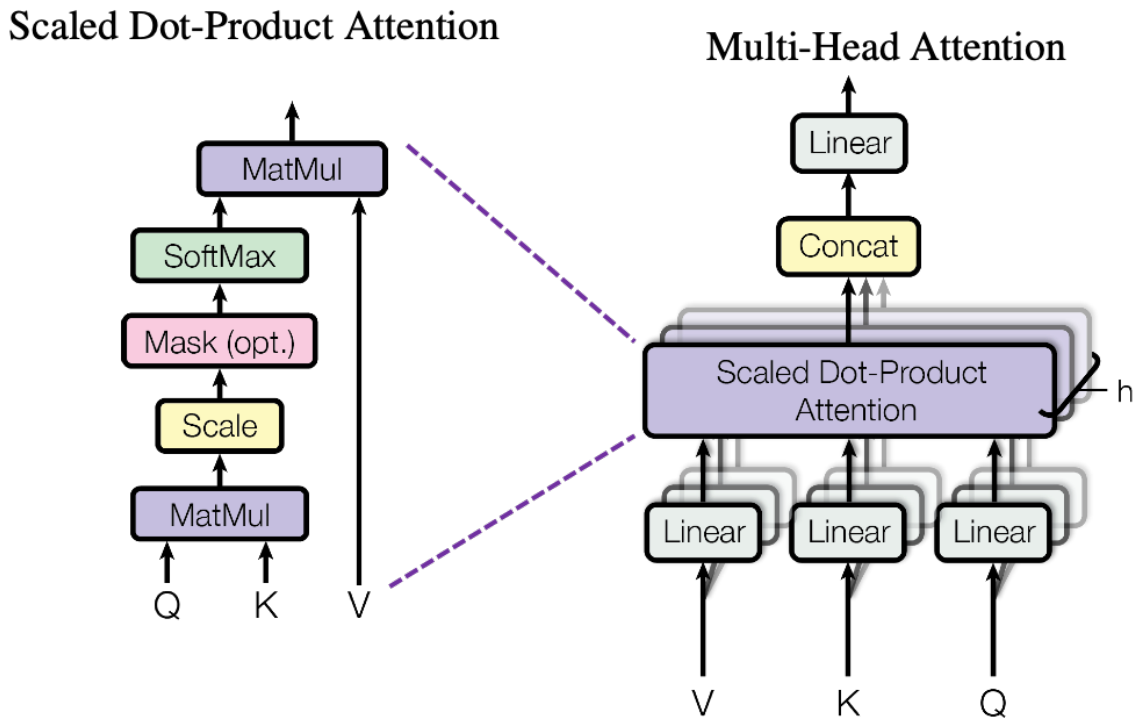

Figure 2.6: (left) Scaled Dot-Product Attention. (right) Multi-head Attention consists of several attention layers running in parallel. Figure taken from Vaswani et al. (2017)

.

2.2.2. Pretraining

Pretraining is integral to what makes BERT so useful for various tasks. Its goal is to make BERT understand the structure and meaning of language by training it on a massive amount of data, and later on, it can be fine-tuned for specific use cases. Pretraining is achieved through two main tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP).

- **Masked Language Modeling** trains the model to predict the randomly masked words in a sentence. During training, 15% of tokens are chosen to be masked. For those, there is an 80% chance to be replaced with the `[MASK]` token, a 10% chance to be a random token, and the remaining 10% of the time remains unchanged. This approach allows the model to obtain bi-directionality, forcing it to capture context better, as it has to consider both the left and right context to predict the masked words accurately.

– **Next Sentence Prediction** is based on predicting whether a pair of sentences follows a logical sequence to let the model understand a relationship between two sentences. It is a binary classification task to determine if sentence `B` is the next sentence that follows `A`. This process enhances BERT's ability to understand longer texts.

3. Data

3.1. Datasets

The data used was accessed from the *hatespeechdata* website, as collated and defined by Vidgen and Derczynski (2021).[1] The focus was on a high-resource language, English, and a few languages with lower coverage, namely French, Portuguese, Indonesian, and Croatian.

A common text preprocessing was done for all datasets, with some more specific processing done for each language. Since most of the sources were comments and tweets, if a letter appeared more than two times in a row, it was reduced to two.

3.1.1. English

The 2021 SemEval task, Pavlopoulos et al. (2021), was focused on finding toxic spans in sentences, so it was primarily a sequence labelling task. Regarding the benchmark on that dataset, the SemEval task differed from the binary classification task of this thesis so the results are not comparable.

However, a dataset was constructed as a subset of re-annotated comments from around 30,000 texts from the Civil Comments dataset by Borkan et al. (2019) that were originally labelled as toxic.[2] Originally, it had the *all-toxic* and *non-toxic* labels with values from a few annotators. To ensure accuracy, we applied a consensus rule that if a comment is considered non-toxic all three annotators must agree. Finally, there were 10543 non-toxic and 5557 toxic comments left, making it the most extensive dataset used. Figure 3.1 shows the distribution of text lengths for this dataset with the median number of words per text being 25, making it the "longest" dataset used.

Due to the dataset not being constructed specifically for the purpose of binary classification, and the label extraction that was performed, there were some inconsistencies with the data. More precisely, instances where the same text received a different overall label. For example, the comment containing only the word "`idiot`" was found 22 times

[1] `https://hatespeechdata.com/`

[2] `https://github.com/ipavlopoulos/toxic_spans`

in the dataset, and while it was labelled as toxic in the majority of those instances, there were still examples of it being marked as non-toxic.

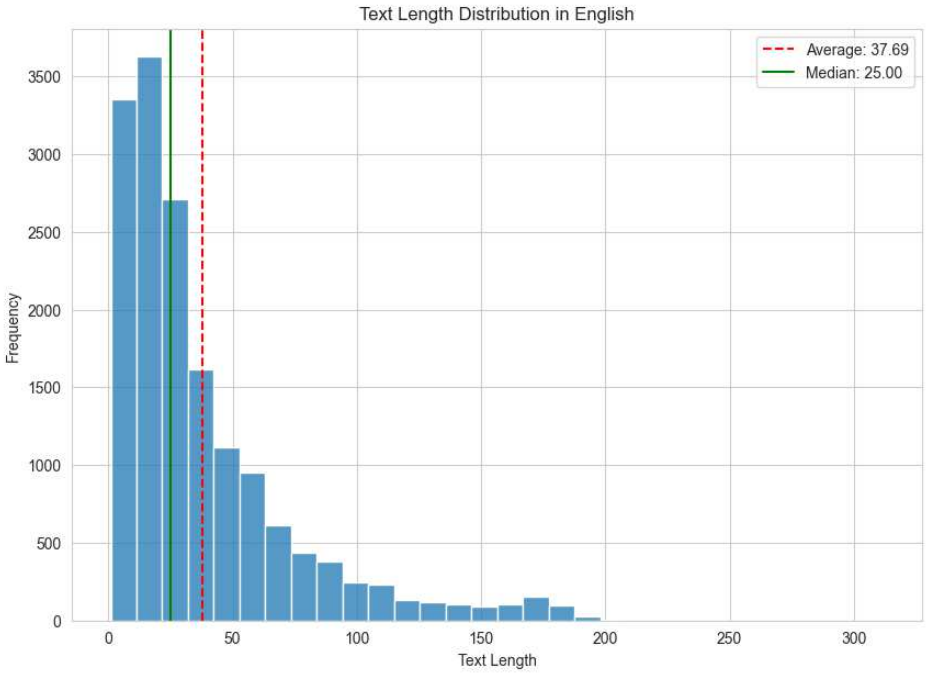

Figure 3.1: Distribution of the number of words per text for the English dataset.

3.1.2. Croatian

The Croatian dataset has been collected by Ljubešić et al. (2018) from the Croatian news site "24 sata".[3] Unlike the other datasets, this one was more substantial. It spans almost ten years of content, and was gathered as an SQL dump of all the comments on the site. The site had one moderator, and after 2016, two moderators deleted hate speech and spam comments. The compiled dataset also included the eight subcategories with the reason why the comments were deleted, as well as the annotating rules. From the manual analysis, only two of them were used to flag the comment as toxic: category 2, which encompasses major direct threats, and category 3 regarding major abuse or derogatory speech. To further ensure the highest possible accuracy of the labels, only comments with dislikes were taken into consideration, which left us with 9462 toxic examples. For the non-toxic comments, since the entire dataset contained over 20 million entries, a portion of it was sampled, making sure that the dislike count on those posts was zero. Lastly, some missing data was removed, along with the same preprocessing applied to all the other datasets. Figure 3.2 shows the word count distribution for the dataset, and Figures 3.3a, and 3.3b show the most common

[3] `https://www.clarin.si/repository/xmlui/handle/11356/1202`

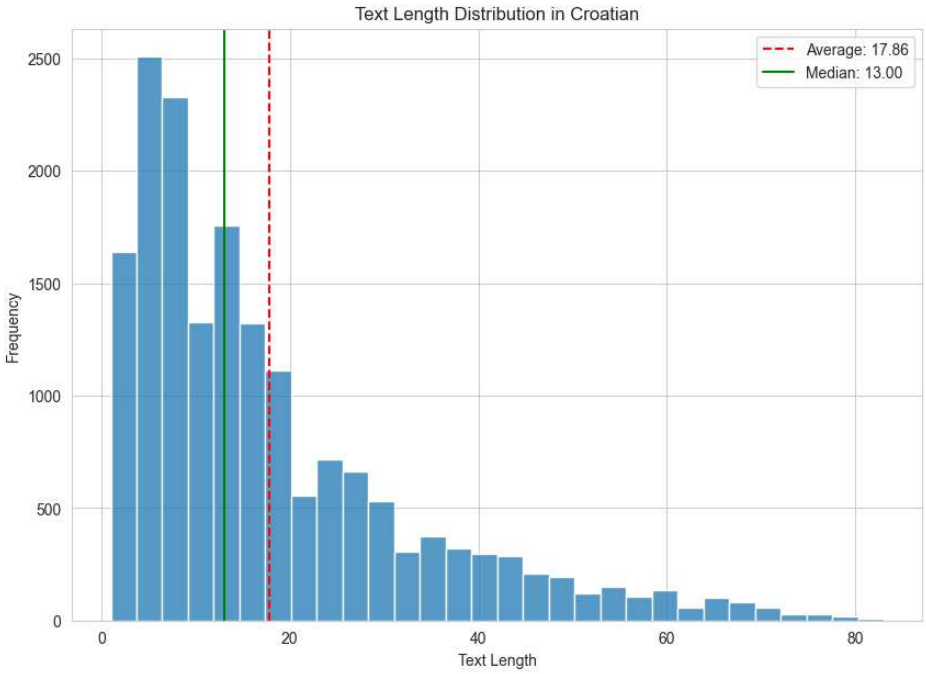

Figure 3.2: Distribution of the number of words per text for the Croatian dataset.

words appearing in the subsets of toxic instances in the English and Croatian datasets, highlighting some familiar profanities and words that often come in sentences with a negative connotation.

(a) English **(b)** Croatian

Figure 3.3: The most common words appearing in the toxic instances of the texts in the English and Croatian dataset.

3.1.3. Indonesian

The Indonesian dataset was created as part of the research paper by Ibrohim and Budi (2019).[4] The data was collected from Twitter and annotated through paid crowdsourc-

[4] https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection

ing. The benchmark accuracy score for hate speech detection for the dataset, according to the publication, is 73.53%.

The original labels came with true or false values for two categories – *hate speech* and *abuse* – along with a mark for the severity of the hate speech. We considered the comment offensive if it was either marked as abusive or hate speech and was ranked at least moderate on the severity scale. Additionally, due to the Twitter-specific elements, some preprocessing was done to remove handles and hashtags. The final value counts were 5829 non-toxic and 3925 toxic instances, and Figure 3.4 shows the word count distribution.

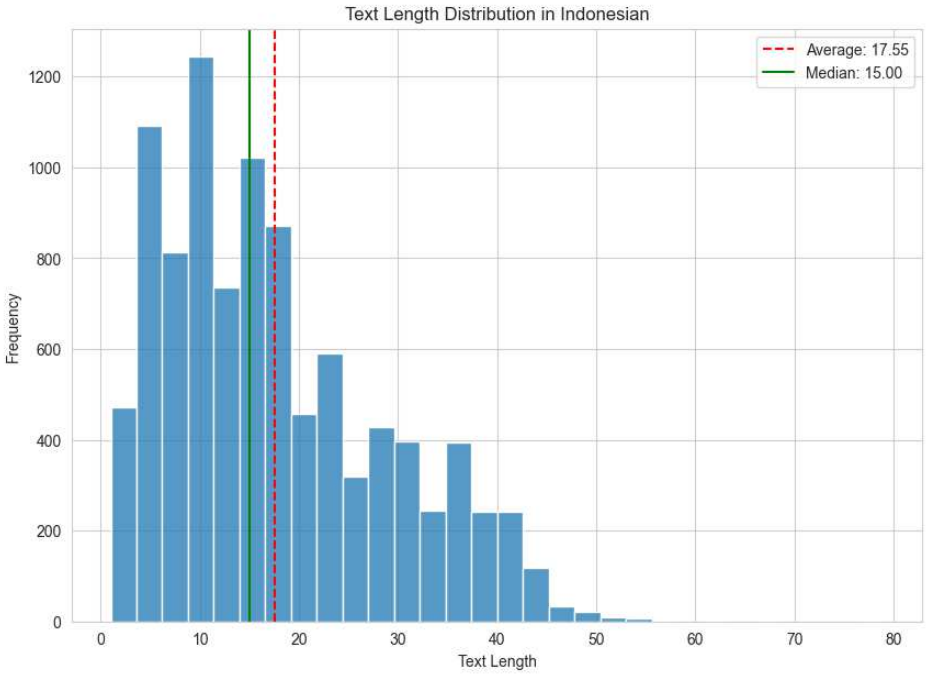

Figure 3.4: Distribution of the number of words per text for the Indonesian dataset.

3.1.4. French

Ousidhoum et al. (2019) collected a hate speech dataset for English, Arabic, and French.[5] The data, similar to the Indonesian dataset, originated from Twitter and was labelled through paid crowd-sourcing. That publication consists of multiple tasks, but on the binary classification task for French, they achieve a macro-F1 of 0.8 and a micro-F1 of 0.69, focusing solely on the directness of hate speech.

The labels in this dataset were more granulated, including *normal*, *offensive*, *abusive*, *fearful*, *hateful*, and *disrespectful*. On top of that, the labels were combined for a total count of 69 different possible labels. For this thesis, a comment was considered as

[5] `https://github.com/HKUST-KnowComp/MLMA_hate_speech`

non-toxic only if it had just the *normal* label. Any other label, except for *fearful* due to ambiguity, classified the text as toxic. Once again, some Twitter-specific preprocessing was required, and finally, 821 non-toxic, and 2943 toxic comments remained.

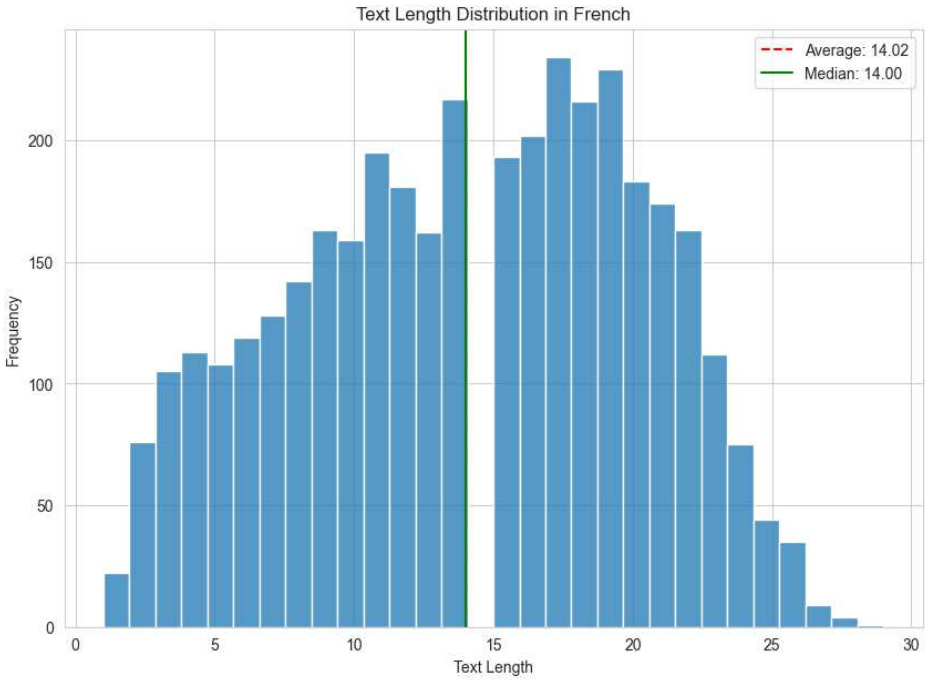

Figure 3.5: Distribution of the number of words per text for the French dataset.

3.1.5. Brazilian Portuguese

The Brazilian Portuguese dataset – `OffComBR-3` – was created by Pelle and Moreira (2017).[6] The data was sourced from comments on a popular Brazilian news portal, "g1", and internally annotated. This specific variant `OffComBR-3` required all three annotators to agree that the comment was offensive for it to be marked as such, in contrast to a more tolerant variant `OffComBR-2`. The publication achieved a maximum F1-score of 0.82.

When it comes to additional dataset-specific processing, nothing was required because the data itself was clean already, and the labels were simply *toxic* and *non-toxic*. However, it was the smallest of the used datasets with 202 non-toxic and 831 toxic comments. Along with having the fewest instances of the datasets used, it also had the lowest average and median word counts of only 13.47, and 11, respectively, as shown in Figure 3.6.

Figure 3.7 shows the final label distributions for all the datasets used. The Croatian dataset remains the largest even after down-sampling, which contrasts with the notion

[6] `https://github.com/rogersdepelle/OffComBR`

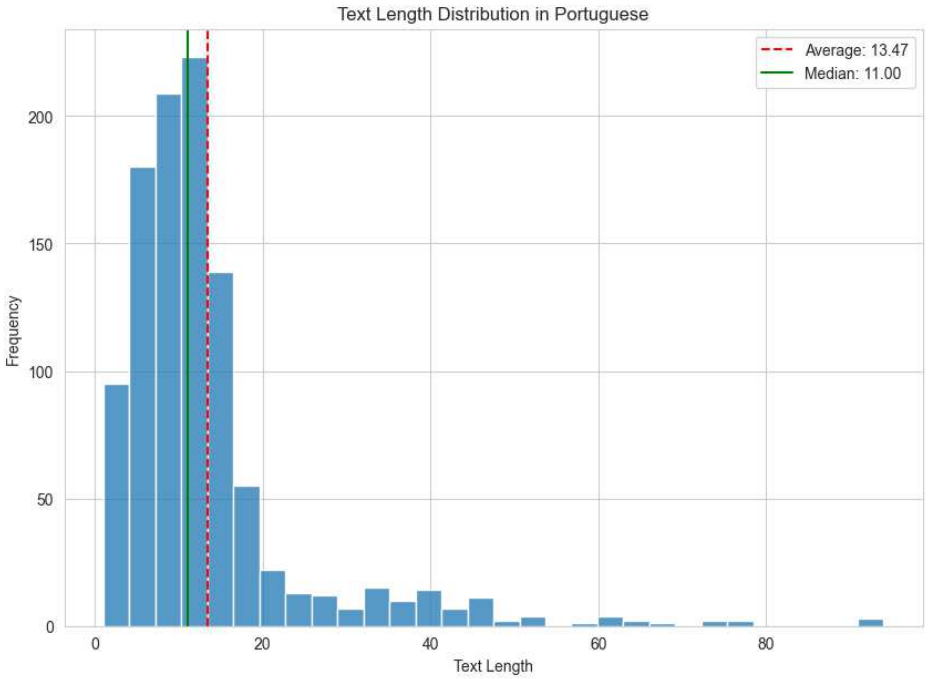

Figure 3.6: Distribution of the number of words per text for the Portuguese dataset.

of Croatian being a low-resource language. However, it is important to note that this dataset is the only one not annotated specifically for research purposes. The labels also had to align with the terms of use of the news site, and the dataset was collected over the longest time span, which may introduce inconsistencies. English and Indonesian fall into the mid-tier in terms of the number of examples for this thesis, with English still being almost twice as large. The French dataset is the only one where toxic comments are the majority class, while Portuguese is by far the smallest with just over 1000 examples.

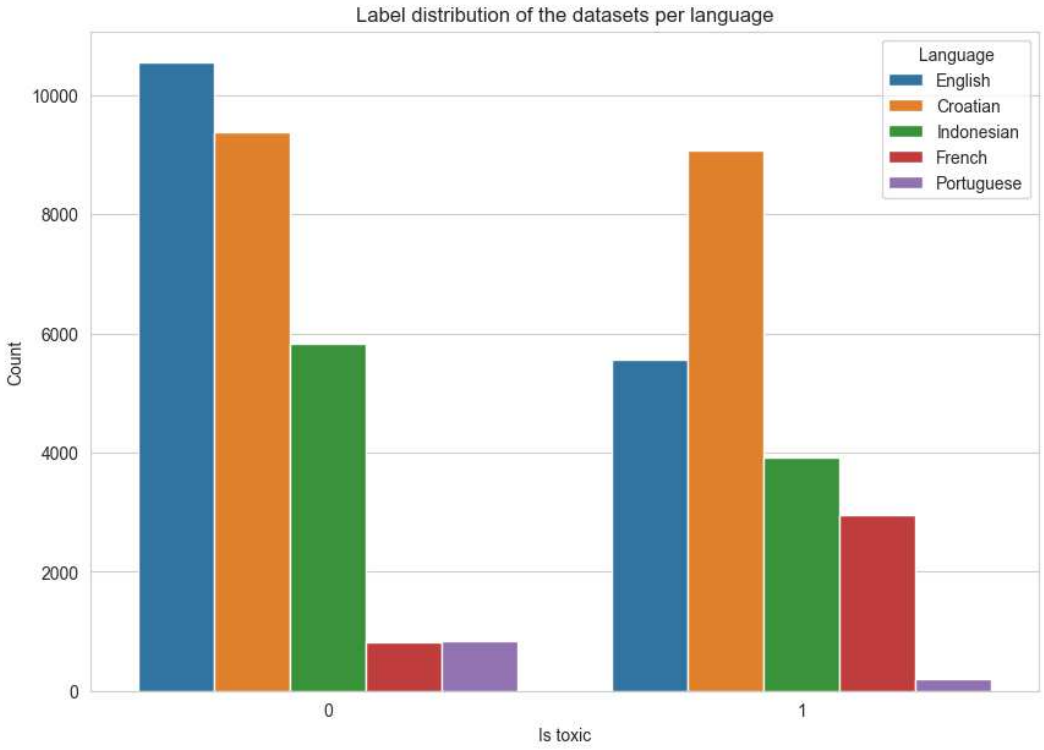

Figure 3.7: Label distributions for all datasets. Class 0 is for the non-toxic, and 1 denotes the toxic comments.

4. Methods

4.1. Baseline

The baseline was chosen as a combination of Term Frequency—Inverse Document Frequency (TF-IDF) for the vectorisation and support vectors for the classification method.

TF-IDF is a statistical method that measures the relevancy of the word within a text relative to a collection of texts. So, if a word appears many times in one document and rarely in others, it is considered to be relevant to that document. It is a way of assigning numerical values to textual data. The implementation used was the one by Pedregosa et al. (2011).

Support Vectors are critical data points that determine the position and orientation of a hyperplane. Support Vector Classification (SVC) aims to find the optimal hyperplane in high-dimensional space to use as a decision boundary that best separates the data. The decision boundary is the equidistant line from the support vectors with different labels. We used the SVC implementation from `scikit-learn` with the Radial Basis Function as the kernel.

4.2. Pretrained Language Models

Large Language Models (LLMs) have significantly evolved over the years, quickly becoming a staple in natural language processing (NLP). These models are often based on transformer architectures introduced by Vaswani et al. (2017) and are signified by their size and capacity to perform complex tasks. Pretrained Language Models (PLMs) is a broader term indicating that a model has been pretrained on a large dataset but can still be fine-tuned for a more specific purpose. The first of these significant models, and the basis for the models used in this thesis, was the Bidirectional Encoder Representations from Transformers (BERT) model presented by Devlin et al. (2018). This model quickly became very popular because it unlocked a new, powerful baseline for researchers. It has been further improved over the years, both in size and in the amount

of training data used. Brown et al. (2020) demonstrate that PLMs are task-agnostic and can perform well with just a small amount of task-specific training data.

There are many advantages to using large language models, such as their ability to understand text in a more nuanced, contextually aware way than previous state-of-the-art solutions. They can be applied to a wide range of tasks, including translation, summarization, and question-answering. They are generally pre-trained on large datasets enabling them to learn a broader range of language patterns. Due to their computationally expensive training, a popular approach is fine-tuning an existing PLM for a specific task. This allows them to work better for certain domains without having to fully retrain them, which enables local deployment and training.

4.2.1. mBERT

mBERT, or multilingual BERT, is a transformer-based model also introduced by Devlin et al. (2018). It was simultaneously trained on data in 104 languages, enabling it to encode the knowledge of all those languages simultaneously. This capability opens up the option of cross-lingual transfer learning, allowing mBERT to apply knowledge contained in the models and data from languages with ample available resources and use it to solve tasks in another, lower-resource language.[1]

The languages chosen were those with the largest Wikipedias at the time of training.[2] Within these languages, there is an inherent bias towards high-resource languages, due to the inevitable skew in the amount of training data. Libovický et al. (2019) found that mBERT consists of both the language-dependent and the language-independent components, demonstrating that transfer learning is easier within language families, but not impossible for dissimilar languages. The mBERT model used for the thesis was the Hugging Face `google-bert/bert-base-multilingual-cased`.

4.2.2. XLMRoBERTa

XLM-RoBERTa or XLM-R is another transformer-based language model developed by Conneau et al. (2019). It was designed to improve upon the cross-lingual language understanding capabilities of mBERT, outperforming it on many cross-lingual benchmarks, and also doing well on low-resource languages. It was pre-trained in 100 languages.

They introduced the concept of the *curse of mulilinguality*, which describes the

[1] `https://paperswithcode.com/task/cross-lingual-transfer`

[2] `https://github.com/google-research/bert/blob/master/multilingual.md#list-of-languages`

trade-off between the number of languages included in the model and its performance on low-resource languages – the performance grows up until a point. The differences between XLM-R and mBERT are that XLM-R uses a masked language modelling objective and was trained on a larger corpus. That significant difference in the amount of data used to train mBERT and XLM-R is shown in Figure 4.1, particularly highlighting the increase for low-resource languages. The variants used are `FacebookAI/xlm-roberta-base` and `FacebookAI/xlm-roberta-large`.

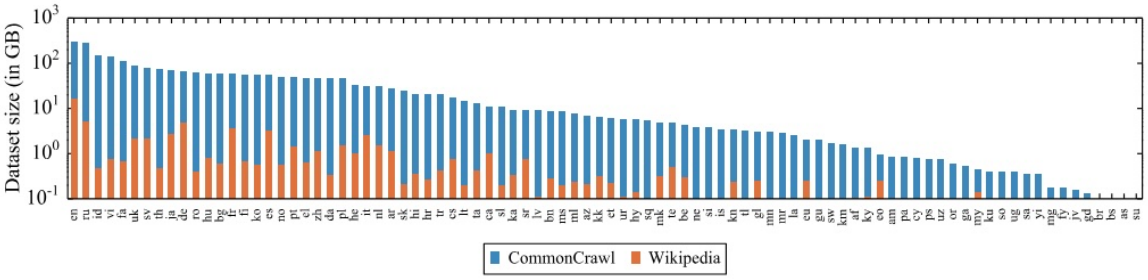

Figure 4.1: Amount of data in GiB (log-scale) for the 88 languages that appear in both the Wiki-100 corpus used for mBERT and XLM-100, and the CC-100 used for XLM-R. CC-100 increases the amount of data by several orders of magnitude, in particular for low-resource languages. Figure taken from (Conneau et al., 2019).

4.2.3. Distilbert

DistilBERT was developed by Hugging Face and introduced by Sanh et al. (2019) as a faster and lighter alternative to BERT. It introduced process of knowledge distillation during the pre-training phase, allowing it to generalise well to many tasks. Knowledge distillation involves training the smaller model (DistilBERT) to reproduce the behaviour of the larger model (BERT) reducing its size by 40% while retaining 97% of its language understanding capabilities. This makes DistilBERT a good option for scenarios where speed is crucial or when computational resources are limited.

The specific variant used – `distilbert/distilbert-base-multilingual-cased` - was trained using mBERT as the teacher model with the same set of languages. Despite its smaller size, it demonstrated comparable performance on many downstream tasks.

4.2.4. GPT-2

The last of the models used was Generative Pre-trained Transformer 2 (GPT-2), developed by Radford et al. (2019) at OpenAI. What differentiates it from the previously mentioned models is that GPT-2 was pretrained solely on a large corpus of English

data called WebText, which was generated by scraping all outbound links from Reddit that received at least three karma.

Unlike BERT, GPT-2 employs a unidirectional architecture that makes it rather effective for tasks such as text generation, which is also its primary purpose. It is able to generate contextually relevant text, and its most popular applications have been in chatbots and content creation. However, it can also be fine-tuned for other tasks, such as sequence classification. It was included due to the popularity of its successor – GPT-4 which was used as the underlying model for ChatGPT. While GPT-4 has not been made publicly available without a subscription, two older variants were tested – `openai-community/gpt2` and the larger `openai-community/gpt2-large`.

4.3. Addressing Data Imbalance

Data imbalance is a common issue in many machine learning tasks, and this study is no exception. It refers to a problem where the number of instances between the classes is not evenly distributed. This imbalance can significantly impact the models' performance because it sees more examples of the majority class during training, which can lead to biased predictions and misleading evaluation. In some cases, like medical diagnosis or anomaly detection tasks, this imbalance, or skew, can be very severe.

To tackle this issue, we used **undersampling**. This technique reduces the number of instances in the majority class to balance the dataset, making it closer to a 50-50 split. This can help improve model performance for the minority class, but it risks losing information from the majority class. Considering the datasets used, this was the logical choice – the labels were not severely imbalanced with the least amount of positives being 33.5% for Portuguese.

Because of this, the main evaluation metrics used were precision, recall, and F1-score, with accuracy being calculated as well for comparison with previous research.

Accuracy is the ratio of correctly labelled examples:

$$accuracy = \frac{correctly\ labelled\ examples}{total\ examples} \tag{4.1}$$

Precision is the ratio of true positive predictions to the total number of positive predictions, measuring the accuracy of the model on the positive label, and is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.2}$$

where TP is the number of true positives and FP is the number of false positives.

Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total number of actual positive instances. It measures the model's ability to correctly find all positive instances and is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.3}$$

where FN is the number of false negatives.

The **F1-score** is the harmonic mean of precision and recall – it considers both false positives and false negatives. The F1-score is defined as:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

The F1-score reaches its best value at 1 (perfect precision and recall) and worst at 0.

4.4. Model Adaptation Techniques

In this thesis, three different approaches were used to adapt the models for the task.

- **Fine-tuning on the Language**:

  The first approach is straightforward: fine-tuning a model for the target language. This enables us to see the performance without additional, task-specific, transfer learning. For example, a pre-trained model is fine-tuned only on Portuguese and then tested on Portuguese as well.

- **Zero-shot Classification**:

  The zero-shot cross-lingual classification is a technique where a model is trained on data in one (or multiple) languages, but is used to classify examples in another language that were not seen during training. This is made possible by the fact that the models used can capture semantic similarities across languages. That ability comes from the models being trained on multiple languages simultaneously, and it falls under the umbrella of transfer learning. For example, a model tuned on English text for the hate-speech detection task could theoretically perform well in Portuguese, which is what we opted to try.

- **Few-shot Classification**:

  The few-shot classification is a middle ground between the two former methods. Essentially, the model is trained on only a few examples of each class. This approach has proven to be useful in scenarios where collecting a large labelled dataset is expensive or challenging. A form of that is training the model on

a task similar to that for which the data exists, or in this case, training it on a language that has a larger dataset available, and fine-tuning it with a few examples of the target language.

In addition to those methods, one more approach was utilised. Since English is predominantly used in both dataset creation and model development, a translation to English was used as the last evaluation technique.

For that, the `deep-translator` library was used, and the model was trained on English, and evaluated on the English translations of the test set in the target language.

5. Experiments and results

5.1. Experimental framework

Training large language models requires extensive under-the-hood optimization to ensure efficiency. This section describes the experimental framework and the libraries used for this purpose.

PyTorch is an open-source deep learning framework widely employed for tasks in computer vision and natural language processing tasks, as detailed by Paszke et al. (2019). It offers GPU acceleration, significantly speeding up PLMs' training time. The framework has a well-developed Python interface, making it a popular choice.

Hugging Face Transformers is another Python library and a repository of models. All of the pre-trained models used were accessed through this library.

For running and training the models, we used **Google Colab** and the resources provided.

5.2. Per language

The experiments were run for all five languages. Firstly, the baseline was calculated to gauge and compare the initial value with the other experiments. Between the mBERT, DistilBERT, XLM-RoBERTa base and large, and GPT-2 base and large, there were six pretrained models to combine with the five possible adaptation techniques. The testing dataset was separated before the experiments were run, so each metric is calculated on the same testing dataset for a particular language.

5.2.1. English

The training for the English model was the only one done without additional modifications besides balancing the dataset since all of the models used were mostly trained in English. Surprisingly, the results have been robust and contain little to no variation. Out of the top five scoring combinations shown in Table 5.1, four of the best scoring models were trained on the entire dataset, except for the best scoring model, possibly

	Model Variant	Adaptation	Acc	P	R	F1
0	tf-idf-svm	/	0.54	0.36	0.47	0.41
1	distilbert-base-multilingual-cased	balanced	0.51	0.51	0.98	**0.67**
2	bert-base-multilingual-cased	/	0.50	0.50	1.00	0.67
3	xlm-roberta-large	/	0.51	0.50	0.98	0.67
4	xlm-roberta-base	/	0.66	0.65	0.68	0.67
5	distilbert-base-multilingual-cased	/	0.66	0.66	0.67	0.66

Table 5.1: Baseline and the best five performing models and adaptation techniques for English.

indicating that the model can identify toxic phrases better if not presented with as much additional data as they are already familiar enough with English.

5.2.2. Croatian

The results for Croatian are similar to those in English and are shown in Table 5.2. The best score on Croatian has been achieved with an XLM-RoBERTa beating the baseline by a few percent. This good performance by the baseline may be because the baseline tokens are trained directly on the dataset, allowing the model to gauge conversational Croatian better. When it comes to the transformer-based models, except for the best score, most of the well-performing models and variants are the DistilBERT and XLM-RoBERTa models. Interestingly, training on almost the entire dataset and using the few-shot adaptation technique with the DistilBERT model produced F1 scores that are not so far behind models fully trained in Croatian. This suggests that the mBERT and DistilBERT could be pretty effective in adapting to the Croatian language, even when data availability is limited. However, the Croatian test dataset was split almost evenly between classes, simplifying the task, and the accuracy barely passed the majority class ratio in six models. This indicates that if a model were to guess the majority class, it would achieve a similar level of accuracy.

5.2.3. Indonesian

The results for Indonesian have fallen into a similar category as those for English and Croatian. With the test dataset consisting of 61% negatives and 39% positives, a higher accuracy could have been achieved by simply predicting the majority class. However, with the best recall being 0.86, as shown in Table 5.3, it is clear that the model has

	Model Variant	Adaptation	Acc	P	R	F1
0	tf-idf-svm	/	0.80	0.86	0.75	0.80
1	xlm-roberta-base	balanced	0.82	0.81	0.84	**0.83**
2	xlm-roberta-base	/	0.79	0.79	0.81	0.80
3	bert-base-multilingual-cased	/	0.79	0.78	0.81	0.80
4	distilbert-base-multilingual-cased	/	0.80	0.80	0.79	0.80
5	distilbert-base-multilingual-cased	balanced	0.78	0.78	0.78	0.78
6	bert-base-multilingual-cased	few-shot	0.65	0.61	0.86	0.71
7	distilbert-base-multilingual-cased	few-shot	0.71	0.71	0.70	0.71
8	gpt2	balanced	0.50	0.50	0.98	0.66
9	bert-base-multilingual-cased	balanced	0.50	0.50	0.96	0.66
10	gpt2	zero-shot	0.50	0.50	0.94	0.65

Table 5.2: Baseline and the best ten performing models and adaptation techniques for Croatian.

been able to capture a fair amount of toxic and hate speech comments. Similarly to Croatian, the baseline outperformed the transformer-based models, although with a higher margin. Indonesian is a language with many dialects, which could also stifle the

	Model Variant	Adaptation	Acc	P	R	F1
0	tf-idf-svm	/	0.86	0.81	0.84	0.83
1	xlm-roberta-base	/	0.85	0.86	0.83	**0.85**
2	xlm-roberta-base	balanced	0.84	0.85	0.84	0.84
3	xlm-roberta-large	balanced	0.83	0.84	0.82	0.83
4	bert-base-multilingual-cased	/	0.81	0.81	0.82	0.82
5	distilbert-base-multilingual-cased	/	0.81	0.81	0.81	0.81
6	bert-base-multilingual-cased	balanced	0.80	0.80	0.80	0.80
7	distilbert-base-multilingual-cased	balanced	0.80	0.79	0.81	0.80
8	xlm-roberta-large	/	0.80	0.81	0.77	0.79
9	distilbert-base-multilingual-cased	few-shot	0.73	0.72	0.74	0.73
10	bert-base-multilingual-cased	few-shot	0.70	0.70	0.69	0.70

Table 5.3: Baseline and the best ten performing models and adaptation techniques for Indonesian.

possibility of capturing all possible tokens with a model pretrained on a different, more formal dataset. Another interesting thing to note is that the GPT-2 models, along with XLM-RoBERTa, have achieved some of the highest F1-scores and accuracy.

5.2.4. French

The evaluation of models on the French dataset revealed some interesting insights. This test dataset had 22% examples of the positive class, making this scenario slightly imbalanced. Notably, the XLM-RoBERTa models, both the base and large versions, training on the entire dataset, were the top performers, achieving the highest F1-score for the French dataset. The mBERT model with the zero-shot approach also achieved a similar F1-score, demonstrating that the task could still be executed in a lower-data scenario. Table 5.4 provides a comparison of the models and their adaptation techniques. Besides the best performers, all of the tested-out models except for GPT-2 models have achieved one of the top ten results in some configurations. All of the top-performing transformer-based models have achieved better results than the baseline.

	Model Variant	Adaptation	Acc	P	R	F1
0	tf-idf-svm	/	0.63	0.89	0.60	0.72
1	xlm-roberta-base	/	0.78	0.77	0.79	**0.78**
2	xlm-roberta-large	/	0.78	0.78	0.78	0.78
3	bert-base-multilingual-cased	zero-shot	0.77	0.77	0.78	0.78
4	xlm-roberta-base	few-shot	0.77	0.77	0.77	0.77
5	distilbert-base-multilingual-cased	few-shot	0.78	0.78	0.76	0.77
6	distilbert-base-multilingual-cased	/	0.77	0.77	0.77	0.77
7	bert-base-multilingual-cased	/	0.76	0.76	0.76	0.76
8	bert-base-multilingual-cased	few-shot	0.76	0.76	0.76	0.76
9	xlm-roberta-base	balanced	0.76	0.78	0.73	0.75
10	bert-base-multilingual-cased	translated	0.73	0.71	0.80	0.75

Table 5.4: Baseline and the best ten performing models and adaptation techniques for French.

5.2.5. Brazilian Portuguese

The results for Portuguese make for an interesting case, particularly when it comes to adaptation techniques. Table 5.5 shows that the DistilBERT model achieved the

best F1-score of 0,78 training on the entire dataset, but a few-shot approach achieves a similar result.

Another thing to note is that plenty of the top-performing combinations have used the zero and few-shot adaptation. Since the dataset on its own was already so small, it is very likely that training on the downstream task benefited from the additional examples provided in English. Finally, it was the only language in which the baseline was significantly weaker than the rest of the models. Since it was the smallest dataset, the training vocabulary likely could not cover many of the testing instances completely.

	Model Variant	Adaptation	Acc	P	R	F1
0	tf-idf-svm	/	0.74	0.45	0.51	0.48
1	distilbert-base-multilingual-cased	/	0.78	0.79	0.78	**0.78**
2	distilbert-base-multilingual-cased	few-shot	0.78	0.78	0.78	0.78
3	bert-base-multilingual-cased	few-shot	0.77	0.77	0.77	0.77
4	xlm-roberta-base	translated	0.76	0.73	0.81	0.77
5	distilbert-base-multilingual-cased	balanced	0.77	0.79	0.75	0.77
6	xlm-roberta-large	translated	0.76	0.76	0.77	0.76
7	xlm-roberta-base	few-shot	0.76	0.76	0.76	0.76
8	bert-base-multilingual-cased	translated	0.76	0.76	0.76	0.76
9	xlm-roberta-base	/	0.76	0.76	0.76	0.76
10	xlm-roberta-large	zero-shot	0.76	0.76	0.76	0.76

Table 5.5: Baseline and the best ten performing models and adaptation techniques for Portuguese.

5.3. Method comparison

Finally, it is important to compare the models and methods used. As shown in Figure 5.1, neither of the larger variants of GPT-2 nor XLM-RoBERTa have outperformed their respective base models in any of the languages. This could lead to the conclusion that bigger is not necessarily better and that using smaller models may lead to equally or more robust results. It also shows that the best model, on average, was DistilBERT, with the baseline following closely behind. When it comes to adaptation techniques, surprisingly, few-shot comes out as a strong contender for training the model on a balanced dataset, as is shown in Figure 5.2. However, the best results have been achieved by training on the entire dataset, and the worst by using zero-shot. Lastly, Figure 5.3

shows that when it comes to the scores per language, the best performances were, on average, achieved on the Portuguese dataset. Still, the best F1 score overall was in Indonesian. Surprisingly, the English dataset had the most considerable variability in results overall.

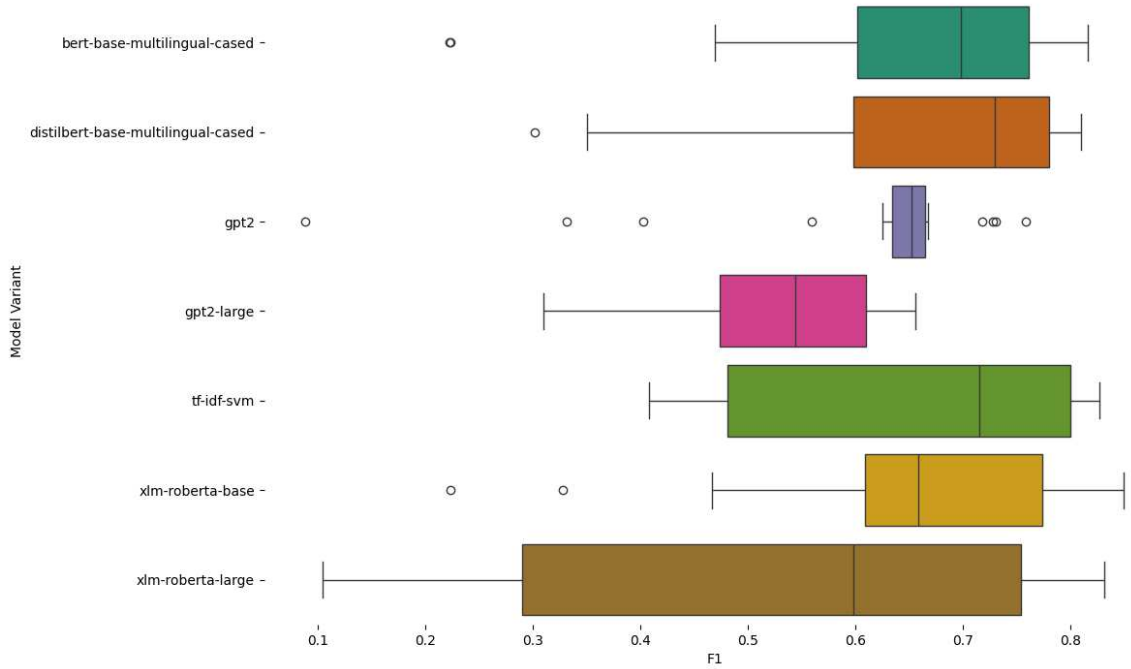

Figure 5.1: This plot shows the ranges of maximum achieved F1-scores for every combination of language and adaptation used per each model tested.

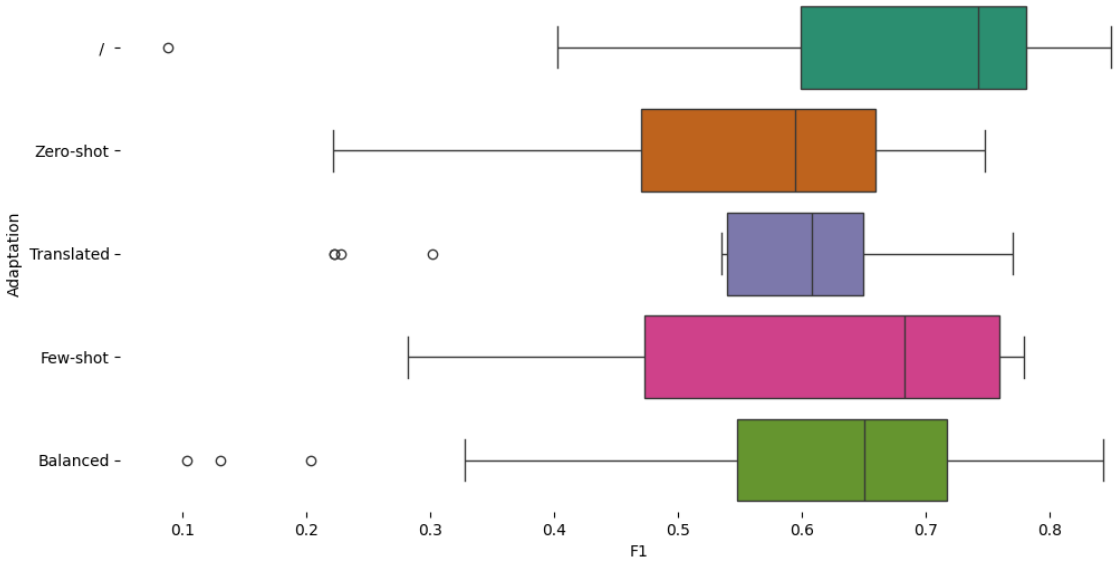

Figure 5.2: This plot shows the ranges of maximum achieved F1-scores for every combination of language and model used per each adaptation technique tested.

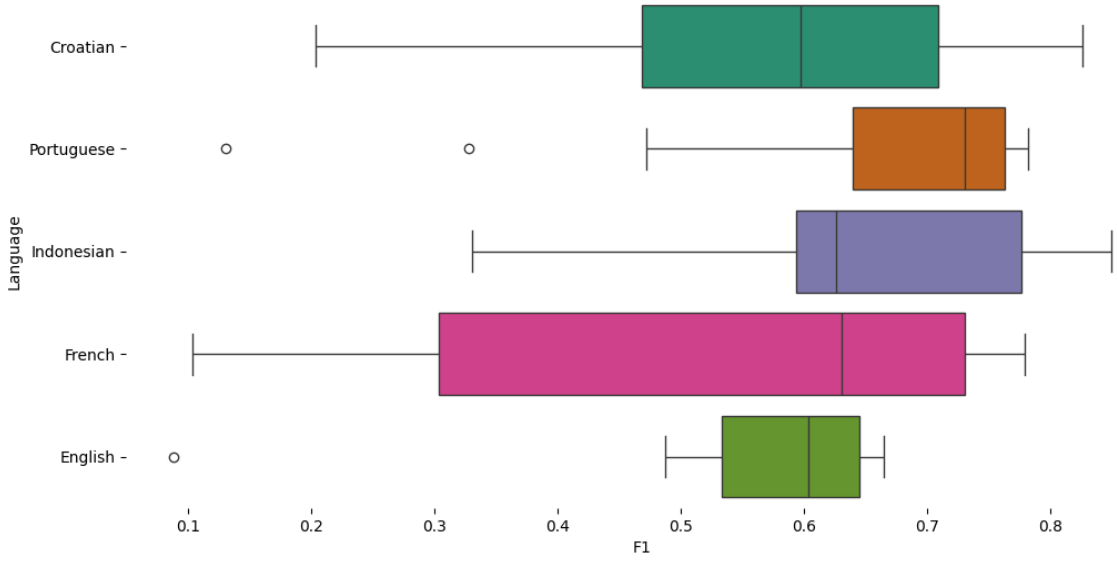

Figure 5.3: This plot shows the ranges of maximum achieved F1-scores for every combination of adaptation and model used per each language tested.

5.4. Analysis

The option of web scraping has enabled the growth in the size of datasets used for NLP. Allowing the automated collection of large volumes of data from online sources enabled easier creation of extensive datasets but also introduced the possibility of a need for more data quality and consistency.

As is the case with the Croatian dataset used for this thesis, the dataset is vast and has been labelled over a more extended period. It would be expensive and difficult to guarantee data quality for a good performance. The model cannot learn properly when the data is significantly noisy – it may overfit on misleading patterns, poorly generalize on unseen data, or not be able to confidently determine a label if there are inconsistencies in the training dataset.

Swayamdipta et al. (2020) propose a method of mapping the model behaviour on individual instances that provides two intuitive metrics – the model's confidence in the label for that example and the variability of that confidence during epochs. The confidence metric is the model's average probability score for the example across epochs. Variability is the standard deviation of the confidence scores across epochs for each instance. With those metrics, such mapping allows a straightforward interpretation of the model's behaviour on the dataset.

The graphed metrics can be grouped into three main categories:

- **Easy-to-learn** instances are the examples for which the model can relatively consistently and confidently determine the correct label. Those examples have a higher confidence and low variability and are usually well-represented in the dataset.
- **Hard-to-learn** instances may be the rarer examples, edge cases or simply more complex for the model to understand. They have low confidence, but despite that, the model's predictions are still stable across epochs, and the variability is low.
- **Ambiguous** examples are the ones with high variability in confidence. In that case, the model predictions fluctuate across epochs, indicating uncertainty. They may occur because of labelling errors, overlapping cases, or noisy data.

Figure 5.4 shows a simplified example of those categories with their positions in the graph.

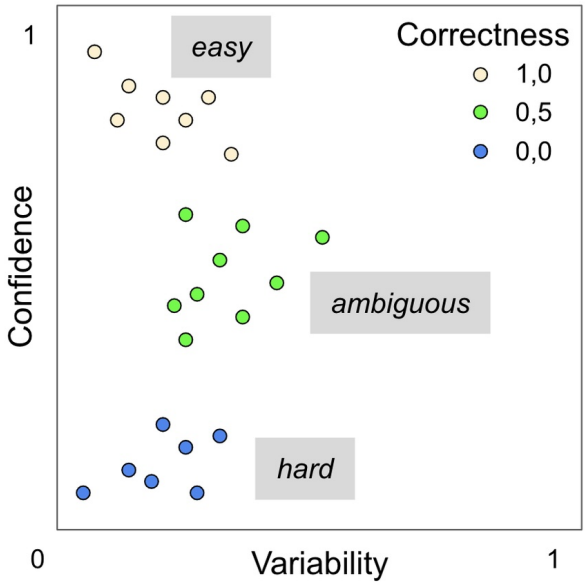

Figure 5.4: Dataset cartography based on training dynamics: average confidence, variability and correctness. Figure taken from the first version of Ponti et al. (2024).

Figure 5.5 shows the training dynamics of the XLM-RoBERTa model on the Croatian dataset.

The values were taken in training during five epochs. While the variability for all examples is not particularly large, the three example groups are visibly separated. The mean confidence for most examples is above 0.5, but overall, it is better for the non-toxic instances, indicating that they are easier to learn. The toxic and profane examples make up most of the harder-to-learn category, with below-average confidence but low variability. The majority of examples fall into the unambiguous group. However, there are still plenty of ambiguous examples, and many of them are either toxic and correctly classified as such or not toxic and incorrectly classified. This could point to the fact that the model leans toward labelling the example as toxic when it is unsure.

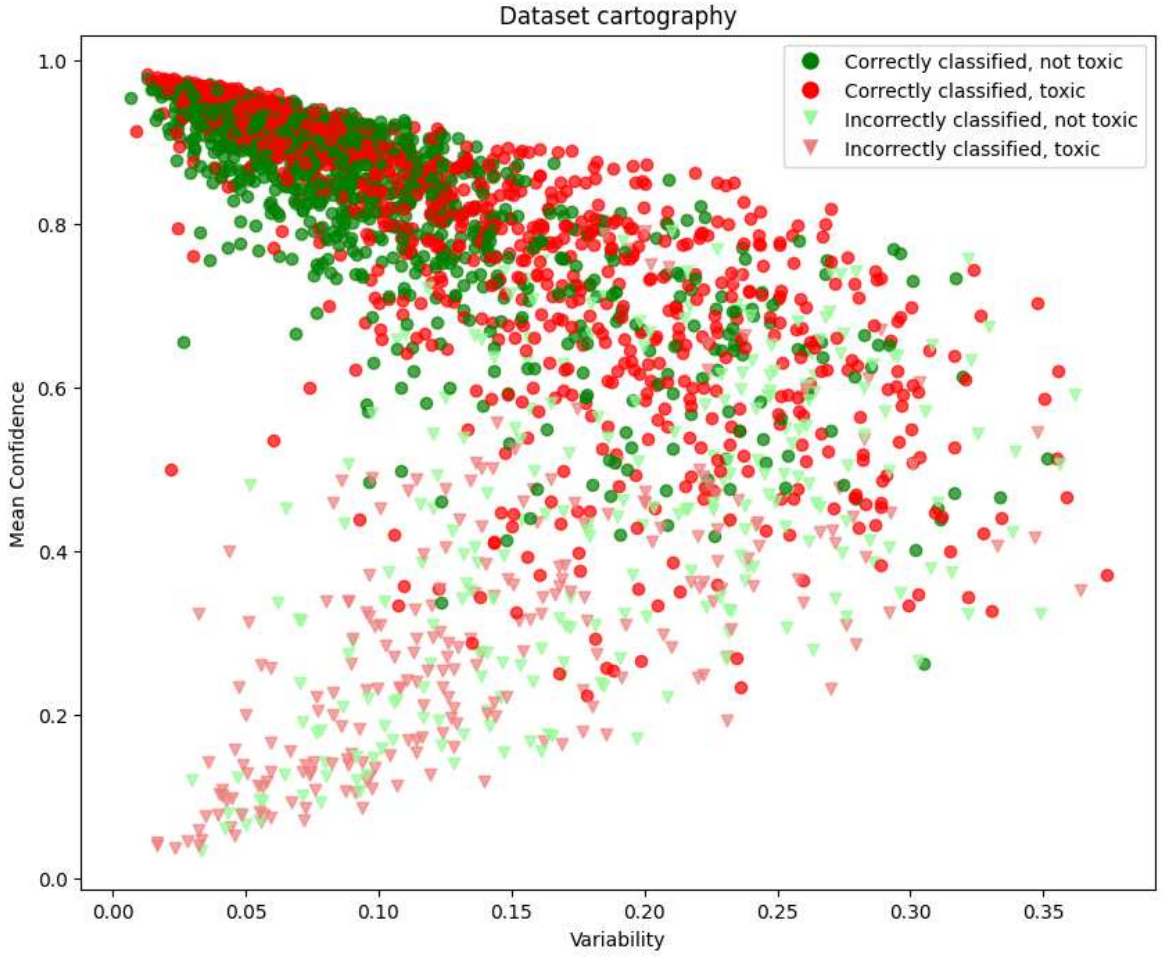

Figure 5.5: Dataset cartography of the Croatian dataset, trained on XLM-RoBERTa in five epochs. The Y-axis denotes the mean confidence of the model in prediction for that instance, the X-axis is the standard deviation of that confidence. Red represents the positive (profane/toxic) examples, and green represents the negative examples.

6. Conclusion

Automated hate speech detection and mitigation is a current and increasingly relevant topic as the internet continues to grow. With the internet, the volume of content produced by users on social media has increased as well, exceeding the capacity for exclusively manual moderation. It is important to try to maintain a respectful environment for all users, which is why many companies have turned to natural language processing and machine learning to identify and possibly remove such content.

Despite the major advancements in machine learning and natural language processing, developing solutions that are effective for various languages is still challenging, even without considering cultural context and the fast evolution of language. Many of the current state-of-the-art models and datasets are predominantly focused on widely spoken languages like English. That prioritization is making it more difficult to gather the necessary resources in other languages for developing an efficient solution. There is not only a lack of annotated high-quality large-scale datasets, but also the specialized models for some languages.

The results of this thesis demonstrate that effective methods exist and that data quantity is not the sole factor in the potential success of a solution. The models used in this thesis performed better on languages that have more resources for pretraining, like Portuguese and French. For the less popular languages, pretrained large language models have not been as impactful. Additional analysis showed that the toxic and profane examples in Croatian have also been more difficult for the model to learn, which could be attributed to the context surrounding them.

One possible approach for future work could be to focus on data quality. A stricter annotation process with a focus in reducing ambiguity and properly representing the relevant categories of toxic and hate speech could produce a smaller, but a higher-quality dataset.

Another approach could explore using ensemble methods where the the outputs of multiple models to are combined to make the final prediction. This could use the strengths of different models, including shallow learning models, to improve the final performance. Additionally, an ensemble of specialized models for each category of

hate speech could be utilized. Overall, there are many options for potential future improvements to this project.

Bibliography

Fatimah Alkomah and Xiaogang Ma. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6), 2022. ISSN 2078-2489. doi: 10.3390/info13060273. URL https://www.mdpi.com/2078-2489/13/6/273.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. Deep learning models for multilingual hate speech detection. *CoRR*, abs/2004.06465, 2020. URL https://arxiv.org/abs/2004.06465.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561, 2019. URL http://arxiv.org/abs/1903.04561.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

Anusha Chhabra and Dinesh Kumar Vishwakarma. A literature survey on multimodal and multilingual automatic hate speech identification. *Multimedia Systems*, 29(3): 1203–1230, Jun 2023. ISSN 1432-1882. doi: 10.1007/s00530-023-01051-8. URL https://doi.org/10.1007/s00530-023-01051-8.

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. URL http://arxiv.org/abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-

training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Batoul Haidar, Maroun Chamoun, and Ahmed Serhrouchni. A Multilingual System for Cyberbullying Detection : Arabic Content Detection using Machine Learning. *Advances in Science, Technology and Engineering Systems Journal*, 2(6):275–284, 2017. doi: 10.25046/aj020634.

Muhammad Okky Ibrohim and Indra Budi. Multi-label hate speech and abusive language detection in Indonesian Twitter. U Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, urednici, *Proceedings of the Third Workshop on Abusive Language Online*, stranice 46–57, Florence, Italy, Kolovoz 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3506. URL https://aclanthology.org/W19-3506.

Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.126232. URL https://www.sciencedirect.com/science/article/pii/S0925231223003557.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. SemEval-2023 task 10: Explainable detection of online sexism. U Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, urednici, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, stranice 2193–2210, Toronto, Canada, Srpanj 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.305. URL https://aclanthology.org/2023.semeval-1.305.

Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. How language-neutral is multilingual bert? *CoRR*, abs/1911.03310, 2019. URL http://arxiv.org/abs/1911.03310.

Nikola Ljubešić, Tomaž Erjavec, and Darja Fišer. Datasets of Slovene and Croatian moderated news comments. U Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, urednici, *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, stranice 124–131, Brussels, Belgium, Listopad 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5116. URL https://aclanthology.org/W18-5116.

Jihyung Moon, Won Ik Cho, and Junbum Lee. BEEP! Korean corpus of online news comments for toxic speech detection. U Lun-Wei Ku and Cheng-Te Li, urednici,

Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, stranice 25–31, Online, Srpanj 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.socialnlp-1.4. URL https://aclanthology.org/2020.socialnlp-1.4.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. *CoRR*, abs/1908.11049, 2019. URL http://arxiv.org/abs/1908.11049.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. U *Advances in Neural Information Processing Systems 32*, stranice 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. SemEval-2021 task 5: Toxic spans detection. U Alexis Palmer, Nathan Schneider, Natalie Schluter, Guy Emerson, Aurelie Herbelot, and Xiaodan Zhu, urednici, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, stranice 59–69, Online, Kolovoz 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.semeval-1.6. URL https://aclanthology.org/2021.semeval-1.6.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Rogers Pelle and Viviane Moreira. Offensive comments in the brazilian web: a dataset and baseline results. 07 2017. doi: 10.5753/brasnam.2017.3260.

Carla Perez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. SemEval-2022 task 4: Patronizing and condescending language detection. U Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan, urednici, *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, stranice 298–307, Seattle, United States, Srpanj 2022. Association for Computational Linguistics. doi: 10.

18653/v1/2022.semeval-1.38. URL https://aclanthology.org/2022.semeval-1.38.

Moacir Antonelli Ponti, Lucas de Angelis Oliveira, Mathias Esteban, Valentina Garcia, Juan Martín Román, and Luis Argerich. Improving data quality with training dynamics of gradient boosting decision trees, 2024. URL https://arxiv.org/abs/2210.11327.

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

Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification for low-resource languages. *CoRR*, abs/2105.05996, 2021. URL https://arxiv.org/abs/2105.05996.

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. Hate-speech and offensive language detection in Roman Urdu. U Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, urednici, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, stranice 2512–2522, Online, Studeni 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.197. URL https://aclanthology.org/2020.emnlp-main.197.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. Multilingual hatecheck: Functional tests for multilingual hate speech detection models, 2022.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics, 2020. URL https://arxiv.org/abs/2009.10795.

Tinkerd. Bert embeddings. https://tinkerd.net/blog/machine-learning/bert-embeddings/, March 26 2023. Understanding BERT | Embeddings.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. Large-scale hate speech detection with cross-domain transfer. U Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis,

urednici, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, stranice 2215–2225, Marseille, France, Lipanj 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.238.

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. URL http://arxiv.org/abs/1706.03762.

Bertie Vidgen and Leon Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32, 12 2021. doi: 10.1371/journal.pone.0243300. URL https://doi.org/10.1371/journal.pone.0243300.

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. URL http://arxiv.org/abs/1903.08983.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çagri Çöltekin. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *CoRR*, abs/2006.07235, 2020. URL https://arxiv.org/abs/2006.07235.

Multilingual offensive and hate speech detection

Abstract

Hate and offensive speech are a present phenomenon in real life and online, affecting many people daily. It is generally considered as an offensive discourse targeted towards an individual or a group based on their characteristics. Due to the vast amount of online content being produced, manual detection of such speech is time-consuming and costly. Therefore, automated detection methods are increasing in popularity. In this thesis, we used natural language processing and deep learning for binary classification of texts collected online in several languages: English, Croatian, Indonesian, French, and Brazilian Portuguese. Multiple models were used, including mBERT, DistilBERT, XLM-RoBERTa, and GPT, as well as additional transfer learning techniques such as zero and few-shot learning. We explored and compared these models and approaches in different languages to identify the ones with the highest performance. The models were primarily evaluated based on the F1 score, a measure considering precision and recall, and compared to similar existing research.

Keywords: Hate speech, offensive speech, classification, natural language processing, large language models

Višejezično otkrivanje govora mržnje i uvredljivog govora

Sažetak

Govor mržnje i uvredljiv govor su fenomeni prisutni u stvarnom životu i na internetu, te svakodnevno pogađaju mnoge ljude. Takvim govorom se općenito smatraju zlonamjerni izrazi usmjereni prema pojedincu ili grupi na temelju njihovih karakteristika. Zbog velike količine sadržaja koji nastaje na internetu, ručno otkrivanje takvog govora je dugotrajno i skupo. Stoga automatizirane metode moderiranja postaju sve popularnije. U ovom se radu koriste obrada prirodnog jezika i duboko učenje za binarnu klasifikaciju tekstova prikupljenih s interneta na nekoliko jezika: engleskom, hrvatskom, indonezijskom, francuskom i brazilskom portugalskom. Korišteno je više modela, uključujući mBERT, DistilBERT, XLM-RoBERTa i GPT, kao i dodatne tehnike prijenosa učenja poput učenja iz nekoliko ili niti jednog primjera. Istražili smo i usporedili ove modele i pristupe na različitim jezicima kako bismo identificirali one s najboljom izvedbom. Modeli su prvenstveno ocijenjeni prema F1 mjeri, koja uzima u obzir preciznost i odziv, te su uspoređeni sa sličnim postojećim istraživanjima.

Ključne riječi: Govor mržnje, uvredljiv govor, klasifikacija, obrada prirodnog jezika, veliki jezični modeli