CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Detecting Social Media Hate Speech Surrounding Refugees Using State-of-the-Art Deep Learning Methods

Master Thesis

CDSCO4001E - Contract No. 22784

MSc in Business Administration and Data Science

Authors:
Frederik Gaasdal Jensen, Student No. 141628
Henry Alexander Stoll, Student No. 141636

Supervisor: Raghava Rao Mukkamala
Co-Supervisor: Sippo Rossi

Submission Date: May 16, 2022
120 pages and 234,675 characters

Disclaimer:
This paper contains strong language that may be considered offensive, profane, or vulgar by some readers.
The writers of this thesis do not condone this type of language, it is solely included for the educational value on the topic.

ABSTRACT

The rise of hate speech on social media is a major cultural threat, as social media platforms are being used to shape the opinions of many people. With an increase of anti-refugee rhetoric in various parts of the world, online hate speech against refugees is becoming a cause of concern for the United Nations (UN), as it has been directly linked to acts of violence and atrocity crimes. Yet, very little research has been conducted regarding detection and analysis of hate speech in the context of refugees. Therefore, this thesis aims at answering the research question "*How can hate speech against refugees on social media platforms be detected and measured using Natural Language Processing methods?*". This work shows that deep learning models can be used successfully to classify hate speech on social media in the context of international refugee crises by relying on transformer-based architectures, leveraging a combination of 12 annotated hate speech datasets from various contexts. The best performing model achieves a macro F1-score of 81.0% on in-domain test data and an accuracy of 73.5% on a refugee-related tweets dataset. Moreover, the model exhibits a solid performance on HATECHECK, a suite of functional test for online hate speech, especially for targeted groups such as refugees and immigrants. By applying the best performing model to English Twitter posts, hate speech levels between 2% and 50% were measured of the datasets surrounding five international refugee crises. Tokens such as "*refugees*", "*refugee*", "*igrants*", and "*immigrants*" from the refugee-related data were among the most influential features when predicting the hate speech class. All the results were validated by UNHCR, the UN refugee agency, and lay the foundation for creating a comprehensive system to measure, monitor, and analyze hate speech against refugees as part of the UN strategy and plan of action on hate speech.

Keywords: Natural Language Processing, Deep Learning, Transformers, RoBERTa, Hate Speech, Refugees, UNHCR, Social Media, Feature Importance

ACKNOWLEDGMENTS

First of all, we would like to take the opportunity to thank our supervisor, Raghava Rao Mukkamala, and co-supervisor, Sippo Rossi, for their dedication and indispensable support throughout the thesis process. The ongoing discussions and feedback that we have received during the process have been much appreciated. We would also like to thank Kwabena Adwabour, Patricia Fabi, and Yawen Tan from the Office of the United Nations High Commissioner for Refugees (UNHCR) for their inputs and feedback, which have contributed to the shape of the thesis. Besides that, we would also like to thank Milos Kovacevic from Copenhagen Business School's Library, Faculty, and Study Services, who have supported us regarding the provisioning of additional computational resources from UCloud. Lastly, we would also like to thank our families for supporting and believing in us throughout the whole process.

CONTENTS

ACRONYMS

BERT Bidirectional Encoder Representations from Transformers.

CRISP-DM Cross Industry Standard Process for Data Mining.

DistilBERT A distilled version of BERT.

HateBERT Re-trained BERT with abusive text data.

IDP Internally Displaced People.
IRR Incidence Rate Ratio.

Linear SVM Linear Support Vector Machine.
LSTM Long Short-Term Memory Neural Network.

ML Machine Learning.
MLM Masked Language Model.

NGO Non-Governmental Organisation.
NLP Natural Language Processing.

OOV Out-of-Vocabulary.

RAL-E The Reddit Abusive Language English dataset.
ReLu Rectified Linear Units.
RoBERTa Robustly Optimized BERT Pretraining Approach.

SHAP SHapley Additive exPlanations.
SVM Support Vector Machine.

TF-IDF Term Frequency Inverse Document Frequency.

UN United Nations.
UNHCR United Nations High Commissioner for Refugees.

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

Around the world, powerful voices are aiming to disparage refugees and make them into objects of hatred and malice. According to the Office of the United Nations High Commissioner for Refugees (UNHCR), the global number of people who have been forcibly displaced from their homes was estimated to exceed 84 million in mid-2021, with around 26.6 million people worldwide being refugees. This is a doubling within the last decade. Due to intensifying violence in Afghanistan, Ethiopia, Syria, South Sudan, Myanmar, and other countries, this number is expected to continue to grow. (UNHCR, 2021).

In the public perception, refugees, migrants, and asylum-seekers are often grouped together and characterized as a highly mobile and predatory type of "foreigners". The hostility towards refugees is therefore frequently the outcome of narratives and attitudes centered around a fear of the outsider, which is mostly based on ethnicity, color, religion, language, and other identity-related factors. Not only can this affect public and political views regarding refugees, which creates barriers in advocacy, fundraising, and lobbying on behalf of refugees, but it can also result in violence and persecution of refugees. Because of this, any kind of communication that attacks or is intended to be derogatory or to humiliate a person or group based on identity-related characteristics, is therefore hate speech. For refugees, this gets especially dangerous when it is combined with explicit incitement for violence, which may lead to acts of terrorism or atrocity crimes (UNHCR, 2019).

Social media plays a big role in this. Even though it can also be used with good intentions by refugee advocates, often these platforms allow the spread of extreme opinions through so-called "echo chambers". These reinforce the existing views of its members through repetition inside a closed system, without encountering opposing views. Because of that, social media can serve as a propagation mechanism for violent acts, as exposure to this content

can push perpetrators beyond just reading to violent actions. In the past, Facebook posts have been linked directly to anti-refugee incidents, including violent crimes like arson of refugee shelters (Müller and Schwarz, 2021). Additionally, in 2018 the UN's fact-finding Mission on the Rohingya genocide cited Facebook as one of the major contributing factors. More than 700,000 Rohingya Muslims in Myanmar were forced to flee amid deadly military crackdowns. Hate speech was amplified as the platform failed to take down inflammatory posts, some which were able to be directly linked to violence (UNHCR, 2019).

1.1 PROBLEM FORMULATION

UNHCR is addressing hate speech against refugees as part of the United Nations' strategy and plan of action regarding hate speech. This is an UN-wide initiative with thirteen commitments, where UNHCR is tasked to among other things contribute towards commitments 1, 2 and 6: "*Monitoring and analysing hate speech*", "*Addressing root causes, drivers and actors of hate speech*", and "*Using technology*" (United Nations, 2020). Currently, from the organization's perspective, the key challenges are associated with analyzing the massive amounts of posts that are generated in the context of refugees on social media. This requires manual detection of anti-refugee sentiment, therefore it is only possible using a sub-sample of data for each refugee crisis.

Automatically and accurately detecting hate speech from social media remains a remarkably difficult task due to the various definitions of hate speech (Waseem and Hovy, 2016). Recent advances in the field of Natural Language Processing (NLP), in the form of deep learning-based transformer architectures, make it possible to classify large amounts of textual data (Vaswani et al., 2017) using the context of a sample. Yet, very little research has been conducted regarding automatic detection and analysis of hate speech in the context of refugees.

1.1.1 Superior Scope and Topic Delimitation

Combining these points results in a superior scope of creating a system to automatically detect hate speech against refugees on all social media platforms in all languages. Because only few social media platforms allow researchers to access their data, and the increased complexity of multi-language NLP methods, this thesis is limited to preexisting datasets in the English language. However, this project design allows for a contribution towards the superior goal given the time and resource constraints. As a result, this work builds the foundation for creating such an automatic detection system. As part of a collaboration with UNHCR, a methodology that applies state-of-the-art NLP methods is proposed to detect hate and offensive speech using data collected from the abusive language detection research field. Furthermore, the resulting model is validated for use in the context of refugees before it is applied to classify large amounts of textual social media posts from Twitter for five different international refugee crises.

1.1.2 Research Questions

Based on the problem formulation and topic delimitation, this thesis aims at answering the following research question (*RQ*).

> *RQ: How can hate speech against refugees on social media platforms be detected and measured using Natural Language Processing methods?*

Sub-questions are formulated below to guide the answering of the research question.

- *Q1: How well can deep learning models help to measure hate and offensive speech in general and for text surrounding refugees?*
- *Q2: Which features are important for identifying hate and offensive speech?*
- *Q3: How does hate and offensive speech change over time in the context of specific international refugee crises?*

1.2 THESIS OUTLINE

To answer the research and supportive questions stated above, the thesis is structured into multiple chapters. In Chapter 2, the conceptual framework is introduced, which establishes a general understanding of the definition of hate speech and refugees together with explaining the technical foundation of a bidirectional transformer architecture. Afterwards, a review of the related literature is provided in Chapter 3. This is followed by Chapter 4, which first describes how the research is conducted by introducing the Research Philosophy and CRISP-DM. Additionally, a methodological explanation of the data collection, preparation, modelling, and evaluation steps are included as well. In Chapter 5, the results of the experiments that were outlined in the methodology are presented, which is followed by a discussion of the results in Chapter 6. Besides that, the discussion also includes recommendations, implications, limitations, and future work. To summarize the thesis, a conclusion is provided in Chapter 7.

2 CONCEPTUAL FRAMEWORK

2.1 HATE SPEECH

Hate speech is associated with multiple definitions depending on whether it is used in an academic, dictionary, legal, or governmental context. To be able to detect hate speech in online social forums, it is crucial to have a clear definition of the term because if a proper definition is not in place, it affects the quality of the data, which will be reflected in the performance of a deep learning model. As a result, this section presents a variety of definitions such that a common understanding of the term can be obtained, which makes it possible to establish a clear definition of how "Hate Speech" is understood in this thesis.

Multiple definitions of the term from various sources are available in Table 1, where it can be observed that there are several variations to the definition of hate speech. Some are more specific than others. Poletto et al. (2021) have analyzed the hate speech research field and published an overview paper, where multiple definitions of hate speech have been collected. From this, it is clear that the definitions also vary, even though all of them are used within academia. This emphasizes that the term has a variety of definitions. By further investigating the literature, it can be observed that multiple papers (Schmidt and Wiegand 2017, De Gibert et al. 2018; Basile et al. 2019) refer to the same definition, which is provided by Nockleby (2000). This is considered positive as it both makes it easier to reuse work from others, compare results, and combine datasets from multiple sources.

To annotate a dataset that consists of sentences from a white supremacy forum, De Gibert et al. (2018) used three rules to determine whether a sentence contained hate speech. These rules have been constructed based on a collection of reviewed literature, and to categorize a sentence as hate speech, all of the three rules have to be in place. In the paper, the following example is used to show a sentence that meets the three rules; *"Islam is a*

Definition	Source
Public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation.	Cambridge University Press (2022)
Use of a sexist or racial slur, attacks a minority, promotes hate speech or violent crime, blatantly misrepresents truth, shows support of problematic hashtags, defends xenophobia or sexism, or contains a screen name that is offensive.	Waseem and Hovy (2016)
Language that is used to express hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group.	Davidson et al. (2017)
Any communication that disparages a target group of people based on some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.	Nockleby (2000)
1) There must be a deliberate attack, 2) directed towards a specific group of people, and 3) motivated by aspects of the group's identity.	De Gibert et al. (2018)
Any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factors.	United Nations (2019)

Table 1: Comparison of different hate speech definitions from literature

false religion however unlike some other false religions it is crude and appeals to crude people such as arabs" (De Gibert et al., 2018). Besides looking at the definitions used in academia, it is also relevant to look at definitions at an intergovernmental level because the main focus of this thesis is on hate speech targeted at refugees. As a result, the definition from United Nations (2019) is included in the table as well. Based on the presented definitions in Table 1, it is possible to state that the definitions all share the same idea of hate speech, but some of them are more specific than others. Since this study focuses on detecting hate speech surrounding refugees, the definition from United Nations (2019) will be used as a general understanding of hate speech.

SIMILAR RELATED TERMS According to Poletto et al. (2021), it is difficult to distinguish hate speech from related terms such as aggressiveness, offensiveness, toxicity, and abusiveness. The reason is that these terms often overlap and contain fuzzy boundaries, therefore, there is a high chance of experiencing strong subjective interpretations of the terms. The separation between such terms from hate speech acts as a key challenge within automatized hate speech detection (Davidson et al., 2017). In a study by Davidson et al. (2017), it was found that previous studies identified the problem of confusing related terms with hate speech. Even though this problem was identified, many studies are still mixing offen-

sive language with hate speech. Because of the multiple definitions and interpretations of the same terms, it makes it difficult to reuse data that has been collected from other studies because a deep learning model will have a difficult time distinguishing between multiple definitions (Fortuna et al., 2020). As a result, it is important to be careful when choosing previously annotated datasets. Also, regarding the compatibility of datasets, it is important to emphasize that the type of words that are used to express hate heavily depend on the targeted groups and context (Fortuna et al., 2020). For example, the same wording might not be used to produce hate speech regarding refugees and the LGBT community.

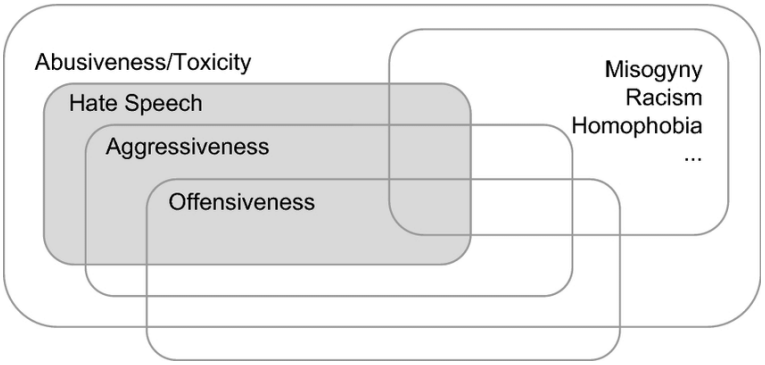

Figure 1: Relations between hate speech and related terms (Poletto et al., 2021)

In Figure 1, the relations between hate speech and other similar terms are visualized. From this, it is clear why it can be difficult to distinguish between these terms since they are overlapping. Regarding the hate speech category, the challenge is to assess when an offensive text can be associated with hate speech because hate speech is not always considered offensive. In Table 2, similar terms such as aggressiveness, offensiveness, and abusiveness are defined to more easily distinguish hate speech from these other related terms. It is important to have a clear understanding of the connection between these closely related terms before annotating data or combining multiple datasets from a variety of sources because it affects the outcome of a deep learning classifier in the end.

2.2 REFUGEES AND RELATED TERMS

In political, media, and public discourse, the terms "refugee", "asylum-seeker", and "migrant" are frequently used interchangeably to describe people who have left their countries and crossed borders. Yet, from a legal perspective, there are crucial distinctions between the terms. These misunderstandings in discussions of asylum and migration are the results

Definition	Source
Aggressiveness The user intention to be aggressive, harmful, or even to incite, in various forms, to violent acts against a given target.	Sanguinetti et al. (2018)
Offensiveness Posts containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.	Zampieri et al. (2019)
Abusiveness Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion.	Founta et al. (2018)

Table 2: Terms related to hate speech

of the public discursively constructing its own definitions (Law et al., 2001). This confusion leads to problems for refugees and asylum-seekers, and for states seeking to respond to mixed movements. To clarify these terms, it will be defined in this section.

The surrounding legal framework protecting the rights of refugees is referred to as "international refugee protection", first asserted in Article 14 of the Universal Declaration of Human Rights, and clearly defined on an international level in the 1951 Convention related to the Status of Refugees, and its 1967 Protocol, also referred to as the Geneva convention. These conventions set forth a universal definition that incorporates the basic rights of refugees. Accordingly, a refugee is defined as a person that

> *"owing to well-founded fear of being persecuted for reasons of race, religion, nationality, membership of a particular social group or political opinion, is outside the country of his nationality and is unable or, owing to such fear, is unwilling to avail himself of the protection of that country; or who, not having a nationality and being outside the country of his former habitual residence as a result of such events, is unable or, owing to such fear, is unwilling to return to it." (UNHCR, 2010)*

The well-founded fear of being persecuted thereby is part of the risks that "classically includes those of persecution, threats to life, freedom or physical integrity arising from armed conflicts, serious public disorder, or different situations of violence" (UNHCR, 2016). If a person is recognized as a refugee, he/she enjoys a specific legal status and is entitled to several important rights and benefits. These include most importantly the right not to

be expelled from the country or returned to situations where their life or freedom comes under threat as well as access to fair and efficient asylum proceedings. States bear the responsibility to guarantee and protect these rights (UNHCR, 2011).

Besides that, it is important to note that under international law, being recognized as a refugee is declaratory, meaning that one does not become a refugee because of recognition, but is recognized because one is a refugee. An asylum-seeker is an individual who is *seeking* international protection. Because of that, not every asylum seeker will be recognized as a refugee, but every recognized refugee is initially an asylum seeker. This status can be confirmed by a state with individualized procedures or by UNHCR. Closely related to refugees are stateless persons and Internally Displaced People (IDP). Stateless persons are not considered to be nationals by any State and include persons whose nationality is not established. Therefore, they are fully covered under the mandate of UNHCR (UNHCR, 2006).

On the other hand, IDP have been forced to leave their homes or place of habitual residence "*particular as a result of or in order to avoid the effects of armed conflicts, situations of generalized violence, violations of human rights or natural or human-made disasters*" (UNHCR, 1998). Similar to refugees, they have many of the same protection needs, but as they have not crossed any internationally recognized borders, they are particularly vulnerable as they remain under the protection of the government, which they have fled (Law et al., 2001). Because of that, they are only covered by UNHCR's mandate in specific circumstances. When focusing on migrants, it can be stated that they are persons who *choose* to leave their countries purely for economic reasons. They are often motivated by the prospect of material improvements in their livelihood or family reunions. The key difference between refugees and migrants, therefore, is the voluntary nature of their migration, as they still enjoy the protection of their home countries, and can always return to their home country without facing persecution or other dangers to their life and well-being (Law et al., 2001). Because of that, economic migrants do not fall within the criteria for refugee status, and are therefore not entitled to benefit from international protection (UNHCR, 2006; Edwards, 2016).

2.3 DEEP LEARNING

To be able to take advantage of deep learning methods for hate speech detection, the concept of deep learning will first be introduced. Deep learning is considered a sub-area of machine learning and aims at implementing parts of the inner-workings of the human brain (IBM Cloud Education, 2020). More specifically, it is based on neural networks, which use input data, weights, and bias parameters to obtain an understanding of data. Besides that, these neural networks have both an input and output layer, which are categorized as visible layers. In between these two layers, multiple connected hidden layers that consist of interconnected neurons are used to model the classification or regression task at hand. This is also where deep learning separates from standard neural networks because when a neural network contains three or more hidden layers, it is considered deep learning (IBM Cloud Education, 2020).

To obtain a prediction, the neural network first performs a forward pass through the network of interconnected layers, where it traverses from the input layer to the output layer. Afterward, the errors are calculated according to a known output, which makes it possible to adjust the weight and bias parameters using an algorithm like Gradient Descent (IBM Cloud Education, 2020). This process is called backpropagation.

The difference between deep learning and machine learning is that deep learning is often better at handling unstructured data such as text and images, whereas machine learning is better suited for structured data. However, machine learning techniques are also able to handle unstructured data, but multiple pre-processing steps have to be implemented as it seeks to obtain a structured format. On the other hand, deep learning reduces the required number of pre-processing steps and automates the feature extraction, which is usually performed manually within machine learning (IBM Cloud Education, 2020).

2.4 TRANSFORMERS

The concept of transformers will now be explained, as this thesis only uses models that are built on top of a transformer architecture. The transformer-based approach introduced by

Vaswani et al. (2017) is a deep network architecture that solely relies on attention mechanisms. Based on multiple experiments, it became evident that this method required less training time together with being more parallelizable. Besides that, Vaswani et al. (2017) outperformed established state-of-the-art models within the area of machine translation. The transformer consists of an encoder and decoder stack, which includes six encoders and decoders stacked on top of each other separately. The rest of the explanation of the architecture is based on a machine translation task.

2.4.1 Encoder

Each of the encoders contain two sub-layers, a multi-head attention layer and a feed forward layer. Besides that, a residual connection is also used for both layers together with a layer normalization. Before iterating through each of the encoders, the input words are converted into a numerical vector representation with an embedding algorithm. This process only takes place in the first encoder starting from the bottom. On the other hand, the rest of the encoders refer to the output of the encoder placed below. To represent the position of each word in a sequence, a positional encoding is added together with the input embeddings.

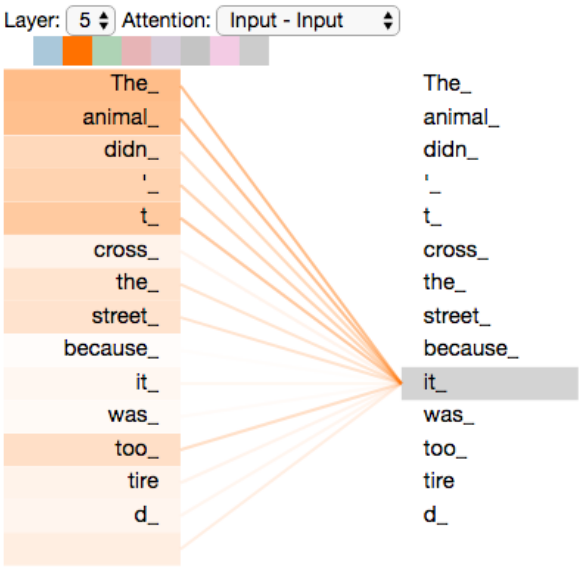

Figure 2: Self-Attention Mechanism. Visualization adapted from Alammar (2018)

MULTI-HEAD ATTENTION When focusing on the multi-head attention layer, it is important to introduce the concept of "self-attention". This is a layer that takes the relationships between each of the words in a sequence into consideration when encoding a specific word. More specifically, when a word is about to be encoded, the connection to the rest of the words in the sequence is considered. This is visualized in Figure 2. The overall process of calculating self-attention includes a query and key-value pairs (these are all vectors), which are created from the encoders' input vectors.

These vectors are used to calculate a softmax score that accounts for the relationship between a single word and all the other words in a sequence. This is performed for each word in a sequence. Afterwards, these softmax scores are used to weight the different value vectors to reduce the impact of irrelevant words, and therefore instead mostly pay attention to highly relevant words. To obtain an output from the attention layer, the weighted value vectors are summed for each word in a sequence. This calculation of the attention is referred to as the "Scaled Dot-Product Attention" by Vaswani et al. (2017) and are performed using matrices as it provides faster processing. A mathematical representation of this calculation for one word is[1]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) * V \tag{1}$$

Besides calculating the attention, Vaswani et al. (2017) also introduces the concept of "multi-head". This means that the attention layer contains several representational sub-spaces. To achieve this, multiple query, key, and value matrices are used, where the attention is calculated in parallel on each of these sets. To reduce the number of output matrices, they are all concatenated into one matrix, which is subsequently multiplied with a weight matrix. The concepts of "Scaled Dot-Product Attention" and "Multi-Head Attention" are summarized in Figure 3.

[1] Q: queries, K: keys, V: values, and d_k accounts for the dimension of a key

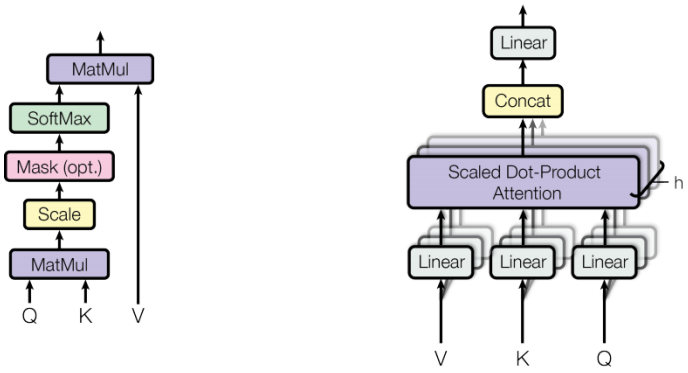

Figure 3: Scaled Dot-Product Attention (LEFT) and Multi-Head Attention (RIGHT). Visualization is adopted from Vaswani et al. (2017)

FEED-FORWARD LAYER Afterwards, the output of the attention layer is used as input in a fully connected feed-forward network that performs two linear transformations and uses Rectified Linear Units (ReLu) as an activation function. Furthermore, one of the performance gains related to a transformer-based approach becomes evident in this layer as each word separately goes through the encoder. This makes it possible to execute the feed-forward layer in parallel because there are not any dependencies associated to each word as in the attention layer.

2.4.2 Decoder

After the input sequence has been processed by the encoder, it outputs a set of attention vectors, which are picked up by each decoder in a multi-head attention layer. This layer is also often referred to as the "encoder-decoder attention layer", and it allows the decoders to focus on relevant parts of the sequence. The difference in the architecture between the encoders and decoders is that the decoders use this additional attention layer, which uses the output from the set of encoders together with a queries matrix that was created based on the output of the layer below as input. Besides that, the decoders also consist of a masked multi-head attention layer, which is a modified version of the multi-head attention layer used by the encoders. This modification masks subsequent positions in a sequence such that the prediction for a specific position only depends on already identified outputs that was predicted before the current position (Vaswani et al., 2017). Like for the encoders, the decoders also consist of a feed-forward layer. A residual connection for each layer is also applied together with a layer normalization as in the encoder architecture.

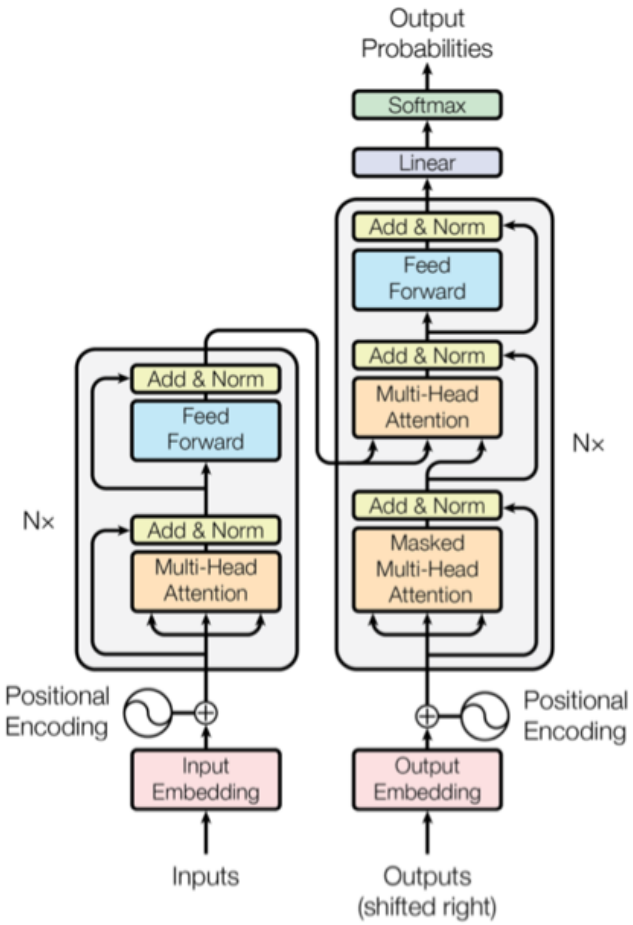

Figure 4: Transformer Architecture from Vaswani et al. (2017)

To obtain a prediction, a fully connected neural network placed on top of the decoder stack receives an output vector from this stack and transforms it into a large vector that consists of logits values. The dimension of this transformed vector depends on the size of the trained vocabulary. Afterwards, a softmax layer is applied on the logits vector to calculate the probability of each word in the trained vocabulary, which makes it possible to choose the word with the highest probability. The visualization of the full transformer architecture is available in Figure 4.

2.4.3 Bidirectional Transformers

The idea of an attention-based transformer has now been explained, which is why it is suitable to introduce bidirectional transformers. Bidirectional Encoder Representations from Transformers (BERT) was introduced by Devlin et al. (2018), and it builds on the attention-based transformer architecture from Vaswani et al. (2017). More specifically, BERT only

relies on the encoding part in a multi-layered fashion. The primary idea with BERT is to use unlabeled textual data to obtain a deep bidirectional representation of the text during a pre-training phase. The bidirectional part comes from the fact that BERT both uses the right and left context of a specific word. Besides pre-training, BERT also allows for a fine-tuning of the model to fit it to specific tasks by adding an additional output layer. The advantages of using such a model is its bidirectional nature, portability, and the inexpensiveness of fine-tuning (Devlin et al., 2018).

The input text to the model is represented using WordPiece embeddings that have a vocabulary size of 30,000 words (Wu et al., 2016). The initial token in a sequence is `[CLS]`, which is a special token that represents the whole sequence. Besides that, sentence pairs are grouped in a sequence, where they are separated using the token `[SEP]`. Additionally, a learned embedding is also used to represent which sentence a token originates from (segment embedding). Before feeding the data into the model, the embeddings for the token, segment, and position are summed to create a representation of each token in the text data. To pre-train BERT, a masked language modelling and next sentence prediction task, which are two unsupervised tasks, are carried out to obtain an understanding language and context (Devlin et al., 2018). These tasks use the BookCorpus (Zhu et al., 2015) and English Wikipedia corpus for training, which consist of 800M and 2,500M words respectively. As a result, Devlin et al. (2018) presents the following two models, $\text{BERT}_{\text{base}}$, which consists of 12 transformer blocks, a hidden size of 768, 12 self-attention heads, and 110M parameters, and $\text{BERT}_{\text{LARGE}}$, which consists of 24 transformer blocks, a hidden size of 1024, 16 self-attention heads, and 340M parameters.

MASKED LANGUAGE MODELLING Inspired by Taylor (1953), the idea is to randomly mask a proportion of the input tokens with the goal to predict the words of the masked tokens by relying on a specific token's context. When the masked tokens' hidden vectors have been obtained, an output layer using a softmax activation function is applied to predict the original word. Cross-entropy loss is used to measure the error (Devlin et al., 2018).

NEXT SENTENCE PREDICTION The purpose of this task is to teach the model about relationships between sentences. More specifically, the model receives a pair of sentences as input, where it has to decide whether the second sentence in the pair actually follows the first one. In the pre-training phase, 50% of the pairs are correct, and the other ones are just random sentences coming from the corpus.

When comparing BERT with an earlier state-of-the-art model such as a Long Short-Term Memory Neural Network (LSTM), the advantages of using BERT is that it provides parallelization, which means that it can handle multiple words simultaneously. Since an LSTM only considers one word at a time, the model requires longer time to train, and as a result, it is more costly. The purpose of BERT is to understand language and context in general such that it can be applied to a variety of downstream tasks. This makes it less expensive to obtain a solid model for a specific task as is not needed to train a model from scratch. Instead, as BERT makes it possible to transfer all the model parameters, a solid model can be obtained by only fine-tuning the model for the specific task (Devlin et al., 2018). Furthermore, the benefit of using BERT is that the model is able to learn deep bidirectional representations of text, whereas prior work has mainly trained models from left-to-right, right-to-left or used a shallow concatenation of these two (Devlin et al., 2018). As a result, BERT makes it possible to obtain more fine-grained representations of text compared to other models. In this thesis, $\text{BERT}_{\text{base}}$ will be referred to as BERT because the $\text{BERT}_{\text{LARGE}}$ is not considered.

3 LITERATURE REVIEW

This chapter provides an exploration of the academic literature and it is not restricted to hate speech detection as there exists relevant work in other closely related areas. More specifically, the technical aspects of the previous work are outlined to create an overview of advantages and disadvantages of a variety of methodologies that have already been tested. Based on this, it is possible to shape the methodology in this thesis.

Recently, the area of text classification has made big advancements due to the increase of available computational power. As a result, it has now become possible to train larger language models with more complicated architectures. This has led to the rise of freely available pre-trained language models that have been trained on massive amounts of data to obtain a general understanding of language at high costs. As a result, it is now possible to take advantage of these pre-trained models by fine-tuning them to suit a specific classification task. This eliminates the cost of training larger models from scratch, and has recently resulted in new state-of-the-art performances in multiple areas of the Natural Language Processing field. In the following sections, different approaches to text classification within the area of abusive language detection spanning from traditional machine learning approaches to state-of-the-art transformer-based approaches are outlined. All the studies that are covered below are summarized in Table 3.

When focusing on hate speech detection regarding refugees, it is limited what has been researched within academia before. Most of the available research provides an analysis of hateful text targeted at refugees but does include a hate speech detection system. One example of this is provided in Arcila-Calderón et al. (2021). As a result, this literature review mostly focuses on hate speech and abusive language detection within other domains as

this thesis is classified as one of the initial works within the area of hate speech detection in the contexts of refugees.

Traditional ML	anti-semitism	template-based strategy, uni-grams	SVM_{linear}	Warner and Hirschberg (2012)
	racism	uni-grams	Naïve Bayes	Kwok and Wang (2013)
	hate speech	Stanford Lexical Parser	Random Forest, SVM, ensemble	Burnap and Williams (2015)
	hate speech, target	n-grams	Logistic Regression	Waseem and Hovy (2016)
	offensive, hate speech	n-grams, tf-idf, POS-tags, Fleisch Reading Ease Score, Flesch-Kincaid Grade Level, sentiment, tweet meta-data	LR, SVM_{linear}	Davidson et al. (2017)
Deep Learning	racist, sexist	task-specific GloVe embeddings	LSTM, Gradient Boosted Decision Tree	Badjatiya et al. (2017)
	hate speech	word2vec, emotion lexicon, character-level and word-level n-grams	BiLSTM with attention layer	Gao and Huang (2018)
Transformer	offensive	$BERT_{base\text{-}uncased}$	fine-tuning	Zhu et al. (2019)
	covert or overt aggression, misogyny	$BERT_{base\text{-}uncased}$	fine-tuning, additional attention layers	Samghabadi et al. (2020)
	hate speech	re-train $BERT_{base\text{-}uncased}$	HateBert	Caselli et al. (2021)

Table 3: Overview of related work by category.

3.1 TRADITIONAL MACHINE LEARNING APPROACHES

One of the earliest studies in the field of hate speech detection was conducted by Warner and Hirschberg (2012), who focused on classifying paragraphs as either anti-semitic or not anti-semitic. The data was obtained both from the American Jewish Congress that provided URLs for websites that previously have been flagged as offensive and Yahoo that offered offensive news group posts. To construct features from the data, Warner and Hirschberg (2012) relied on a template-based strategy from Yarowsky (1994). More specifically, word n-grams, part-of-speech tagged words, and brown clusters (Koo et al., 2008) were used as features. Furthermore, a feature that represented whether some words did occur in a word window of ten words was used. To perform the classification, a Support Vector Machine (SVM) model including a linear kernel was used and evaluated in a 10-fold cross-validation setting. The best F1-score of 63% was achieved using a uni-gram representation (Warner and Hirschberg, 2012).

Kwok and Wang (2013) also used a supervised Machine Learning approach to perform a different type of binary classification of tweets into the two classes racist and nonracist. More specifically, a Naïve Bayes classifier with bag-of-words features was applied, which

was evaluated using a 10-fold cross-validation approach. The input to the model consisted of a balanced dataset, which contained 25k tweets, where stopwords, user mentions, punctuation, and URLs were removed. Besides that, all the characters were transformed into lowercase and spellings of slurs were corrected. The model achieved an average accuracy of 76%, and it was found that since the data was represented with unigrams, the model was not able to consider the context the words appeared in. Furthermore, it is also stated that the racial identity of the person who wrote a tweet, often are taken into consideration by humans when deciding whether a tweet contains racism (Kwok and Wang, 2013).

Burnap and Williams (2015) had a different focus than the two other papers as they wanted to both classify hate speech and the spread of it. They collected tweets in a period of 14 days after the Drummer Lee Rigby was murdered in 2013 in Woolwich to analyze how online hate speech were spreading in case of such an event. The reason for only using a two week window for collecting tweets was based on former research, where it was found that the public interest was rising a lot in a short time-frame following an event before decreasing again (Downs, 1972). The data was collected using the Twitter API and a subset of tweets were randomly chosen to be annotated into two categories, cyber hate and general response. To construct features, the Stanford Lexical Parser and context-free lexical parsing model were used to obtain typed dependencies in the tweets, which represented the syntactic grammatical relationships. Furthermore, the tweets were also represented with a bag-of-words technique using n-grams. Multiple experiments were run using these features to obtain the most optimal feature set. For the preprocessing of the text, each character was transformed into lowercase, most non-alphanumeric characters were removed together with stopwords, and each token was stemmed.

To perform the classification task, a Bayesian Logistic Regression, Random Forest Decision tree, and Support Vector Machine were used together with an ensemble classifier that combined these individual models. More specifically, a voting meta-classifier that relied on maximum probability were used as an ensemble technique. To validate the models, a 10-fold cross-validation method was used. Based on the results, it was found that when combining n-grams of antagonistic and hateful terms together with typed dependencies

represented in n-grams, the best F1-score of 77% was achieved (Burnap and Williams, 2015). The same score was reached in all the models but the ensemble model had a higher recall in some experiments, which is why this was chosen to be the best model.

Besides analyzing the model's ability to detect hate speech, the authors also wanted to determine the spread of such kind of content on Twitter. Burnap and Williams (2015) stated that the number of retweets is highly associated with the number of people being exposed to a specific tweet. This has a negative effect if the content is considered hateful as it increases the risk of being exposed to the wider public. To measure the spread, a zero-inflated negative binomial model was used where the number of retweets was treated as the dependent variable. The statistical predictors in the model were based on a variety of meta-data together with the prediction from the classification model. To represent the strength of causal relationships between each predictor and the dependent variable, the Incidence Rate Ratio (IRR) was used. Based on the results, it was found that tweets that contained hateful/antagonistic content were reducing the IRR by a factor of 0.55 (Burnap and Williams, 2015). This means that the likelihood of spreading hateful tweets in this case to the wider public was reduced. According to Burnap and Williams (2015), decision- and policymakers can use this type of analysis to study fluctuations in the public opinion following a trigger event.

Waseem and Hovy (2016) used a list of multiple criteria for annotating around 16k English tweets, which had been collected using the Twitter API. In an analysis of the collected data, it was found that the majority of the hate speech were produced by men, but since men constitute the largest part of the dataset in general, this cannot be considered conclusive. On the other hand, the findings of Roberts et al. (2013) state that men constitute 87% of perpetrators of hate crimes in relation to Asians, and 75% in relation to African Caribbeans, which supports the insight discovered by Waseem and Hovy (2016). Furthermore, a Logistic Regression model was used to perform the classification task, and to evaluate the model, a 10-fold cross-validation was applied. As features, word and character n-grams of the tweets were used together with information about the gender of the user, and the location of a tweet. Additionally, the total and average length of a tweet were used as well

together with the length of the user description. By applying the Logistic Regression model and multiple combinations of these features, Waseem and Hovy (2016) found that character n-grams together with information about the gender of a user achieved the best F1-score of 74%. Furthermore, the results indicated that using the length and location as features had a negative effect on the performance of the model.

Davidson et al. (2017) used a lexicon-based method to collect data from the Twitter API, which was subsequently annotated into the following three classes, offensive, hate speech or neither. To test the quality of the dataset, experiments using multiple models such as Naïve Bayes, Logistic Regression, Linear Support Vector Machine, Decision Tree, and Random Forest have been performed. Before running the models, several features were constructed. The tweets were stemmed and represented as uni-, bi-, and trigrams with a Term Frequency Inverse Document Frequency (TF-IDF) value. Moreover, the syntactical structure of the tweets were also captured using Part-of-Speech (POS) tags, and to account for the quality of a tweet, the Flesch Reading Ease score were used together with the Flesch-Kincaid Grade Level (Davidson et al., 2017). To represent the sentiment of a tweet, a sentiment lexicon that were specifically created for social media by Hutto and Gilbert (2014) were applied. Furthermore, a count and binary representation were used to account for retweets, URLs, hashtags, and user mentions in the tweets. Lastly, the number of words, syllables, and characters were used as features as well.

The models were tested using a 5-fold cross-validation, which indicated that the Linear SVMand Logistic Regression models outperformed the other models. To better be able to analyze the predicted results, the Logistic Regression model with L2 regularization was used as their final model. A one-versus-rest framework was applied to evaluate the model, where an F1-score of 90% was achieved (Davidson et al., 2017). Acccording to Davidson et al. (2017), earlier studies have relied on a too broad definition of hate speech, which resulted in offensive text being mislabeled as hate speech. This is minimized in the study by Davidson et al. (2017) due to the use of a multi-class framework. As a result, only around 5% of the offensive texts has been mislabeled as hate speech.

3.2 DEEP LEARNING-BASED APPROACHES

As the availability of computational resources and power has increased in recent years, deep learning methods have experienced a rise in popularity. Deep learning has made it possible to handle large quantities of data and construct models with millions of parameters. As a result, NLP in general and the field of hate speech detection have adopted Deep Learning successfully and outperformed traditional Machine Learning methods.

Badjatiya et al. (2017) experimented with multiple deep learning architectures to learn task-specific embeddings that could be used to detect hate speech, more specifically whether a tweet was racist, sexist, or neither. To learn these semantic word embeddings, FastText, Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) Neural Networks were used, and the input was initialized either with GloVe embeddings or random embeddings. The GloVe embeddings that were used to train the task-specific embeddings were trained on a corpus that included 2B tweets, 27B tokens, and a vocabulary size of 1.2M (Pennington et al., 2014). From multiple tests, it was found that the size of the embeddings was less important. To subsequently test the quality of these embeddings, multiple classifiers were used such as a Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosted Decision Tree, and Deep Neural Network. Based on the results of the classifiers, Badjatiya et al. (2017) found that training task-specific embeddings using deep neural networks before using the embeddings in a classifier significantly outperforms methods that already exists. The best performing method was an LSTM that was trained on random initialized embeddings to obtain task-specific embeddings that were used as an input in a Gradient Boosted Decision Tree, which achieved an F1-score of 93% (Badjatiya et al., 2017).

Gao and Huang (2018) highlight the importance of using contextual information to detect hate speech within text. In the study, 1,528 user comments have been collected from Fox News, and these comments came from the discussion thread of 10 different articles. The data includes information about the context such as the comments and their nested structure, user screen names, and the associated news article (Gao and Huang, 2018).

To detect whether a sentence contains hate speech, Gao and Huang (2018) utilized three types of models, Logistic Regression, Neural Network, and an Ensemble model. In the Logistic Regression model, character-level and word-level n-grams were used as features together with additional features extracted from two lexicons (LIWC and NRC emotion lexicon). Furthermore, an LSTM model with three different inputs was used as a neural approach to this specific task. The three inputs consisted of the user comment, username, and the title of the news article, where the news title and user comment were represented with word embeddings pre-trained using word2vec. Besides that, the usernames were represented as a sequence of characters. To perform the modelling, a bidirectional LSTM model was used on the usernames and news titles, whereas a bidirectional LSTM with an attention mechanism was applied on the user comments.

Based on the results of the experiments, the Logistic Regression model with all of the four features performed best with an F1-score of 54% and AUC-score of 0.78 compared to experiments that tested different combinations of the features for that model. When using the titles of the news articles as additional input, the neural model achieved the best F1-score of 55%. However, using the username together with the news title as additional context input, the model obtained the best AUC-score of 0.77. Furthermore, Gao and Huang (2018) combined the prediction from the best neural network with the best logistic regression model to construct an ensemble model, which achieved better scores than when the models were used individually. Using a Max Score Ensemble, gives the best F1-score of 60%, whereas an Average Score Ensemble obtains the best AUC-score, which is 0.80.

Based on the results, Gao and Huang (2018) state that using character-level n-grams in the Logistic Regression model is a powerful way to detect hateful comments that includes misspelled words, Out-of-Vocabulary (OOV) words or capitalized words. Additionally, it is also stated that using an LSTM model with an attention mechanism, makes it possible to identify smaller areas of a long comment that are hateful. Based on the findings in this paper, it can be argued that contextual information is important when detecting hate speech in online forums.

3.3 TRANSFORMER-BASED APPROACHES

Previously, Recurrent-, Gated-, and Long Short-Term Memory Neural Networks achieved state-of-the-art results in sequence modelling, but in recent years these models have been surpassed by transformer-based approaches (Vaswani et al., 2017). More specifically, by Bidirectional Encoder Representations from Transformers (BERT) models (Devlin et al., 2018), which constantly produces state-of-the-art results.

Zhu et al. (2019) used a pre-trained BERT model (Devlin et al., 2018) to determine whether a tweet was offensive or not based on data provided by Zampieri et al. (2019), which contained 13k tweets. The pre-trained BERT model was further fine-tuned and a linear layer for classification of the text sequences was placed on top of the model. Before feeding the input data into the model, the sentences were tokenized using the BERT basic tokenizer, which converts a sentence into tokens, characters into lowercase, performs punctuation splitting, and removes invalid characters. Furthermore, if a word was not in the vocabulary, each word was split into sub-word units using WordPiece tokenization (Wu et al., 2016). To account for the order of the sequences, positional embeddings were used.

The model clearly outperformed the baseline as it reached a macro F1-score of 81%, which resulted in a third place out of 103 submissions for the SemEval-2019 Task 6 (Zhu et al., 2019). Based on the results, it was argued that the model was performing well at handling the minority class of the offensive tweets because the difference between the F1-score and accuracy score is small. An error analysis of the model's misclassifications shows that it has a difficult time understanding the larger context that the words appear in and that it mostly understands trigger words that have a negative connotation (Zhu et al., 2019).

Further research also successfully used a BERT model in different domains. Samghabadi et al. (2020) proposed a BERT model that includes attention on top of it to perform multi-task classification for the shared task, TRAC-2, regarding *Aggression Identification* and *Misogynistic Aggression Identification* (Kumar et al., 2020). The dataset were provided by Bhattacharya et al. (2020), and it consists of text in both Bengali, English, and Hindi. For the

sake of this thesis, the approach regarding the English data will only be considered because the thesis is constrained to the English language. As the shared task contains two sub-tasks, the data had multiple labels. For the *Aggression Identification* task, the following three labels were available: *Not Aggressive*, *Covertly Aggressive*, and *Overtly Aggressive*. Regarding the *Misogynistic Aggression Identification*, the following labels were used: *Gendered* and *Non-gendered*.

To solve the multi-task classification problem, Samghabadi et al. (2020) used a BERT layer to obtain information about the context followed by an attention layer that was proposed by Bahdanau et al. (2014). Additionally, two fully-connected linear layers were used to reduce the dimension. Lastly, to obtain a final prediction for the each sub-task, two separate classification layers were applied. To prepare and tokenize the input data, the BERT tokenizer has been used and pre-trained weights were obtained from $\text{BERT}_{\text{base-uncased}}$ (Devlin et al., 2018). Based on the results, the model achieved a weighted F1-score of 71% for *Aggression Identification* and 86% for *Misogynistic Aggression Identification*, which led to the model being ranked third out of 15 teams for TRAC-2 (Kumar et al., 2020).

As hate speech on social media is a complex phenomenon, it has been proposed that the key obstacles for advances in detection of hate speech and other abusive language phenomena lie with the quality of data (Fortuna et al., 2020). Moreover, annotated hate speech corpora and benchmarks are important resources, given the considerable amount of supervised approaches that have been proposed (Poletto et al., 2021). Especially when comparing the results of different BERT-based models on the OffensEval 2019 dataset (Zampieri et al., 2019), it seems that the choice and therefore quality of training data has the biggest impact on results. This is in line with other research in the field of NLP and hate speech (Liu et al., 2019; Swamy et al., 2019).

Caselli et al. (2021) identified that state-of-the-art BERT models for hate speech detection are often pre-trained for general purpose language understanding tasks, which limit their capabilities within certain domain-specific tasks. Based on previous work regarding the creation of domain-specific models like TweetEval for interpreting Twitter data (Barbieri

et al., 2020), and LEGAL-BERT for the English legal domain (Chalkidis et al., 2020), Caselli et al. (2021) developed HateBERT to provide evidence that further re-training is a viable strategy to obtain domain-specific models.

More specifically, HateBERT is based on the English $\text{BERT}_{\text{base-uncased}}$, but to obtain a domain-specific model, it is re-trained using the Masked Language Model (MLM) objective on RAL-E, a large-scale dataset of social media posts in English from banned communities on the social media site Reddit. These user-created and user-moderated communities were banned for being offensive, abusive, or hateful after the website strengthened its content policies. As result of the re-training, the models language polarity (i.e., offense-, abuse-, and hate-oriented) shifts slightly compared to the original English BERT model. HateBERT outperforms the generic BERT in different abusive language benchmarks with a macro F1-score of 81% in OffensEval 2019 for offensive language detection Zampieri et al. (2019), 77% in AbusEval, a dataset consisting of OffensEval 2019 with an added layer of abusive language annotations (Caselli et al., 2020), and an 52% for the HatEval hate speech benchmark dataset (Basile et al., 2019; Caselli et al., 2021).

4 METHODOLOGY

This chapter presents the different methods that were used to conduct the experiments in this thesis. The Cross Industry Standard Process for Data Mining (CRISP-DM) framework proposed by (Chapman et al., 2000) was used to structure parts of this thesis' workflow.

4.1 RESEARCH FRAMEWORK

In this section, the framework of the research that is conducted in this thesis is outlined using the "Research Onion" proposed by Saunders et al. (2019). More specifically, it consists of multiple stages that are used to conduct research. Therefore, it is applied to develop knowledge in this thesis. The framework is visualized in Figure 5, and consists of the following six stages, research philosophy, research approach, methodological research choice, research strategy, research time horizon, and research techniques and procedures. The process of developing research can be understood as the process of going from one layer to the next, where each step increases the level of detail. To successfully apply this framework, each stage must be iterated properly. In the following sections, each of the stages in the "Research Onion" are covered.

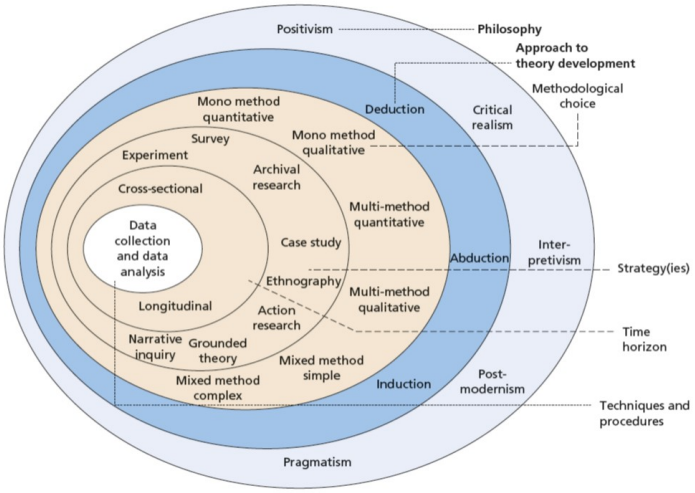

Figure 5: Research Onion from Saunders et al. (2019)

4.1.1 Research Philosophy

Saunders et al. (2019) state that a research philosophy is based on assumptions and beliefs regarding the development of knowledge. These assumptions help with creating a more clear understanding of the research and the methods themselves. Consistency among well-thought assumptions is therefore key to construct a credible philosophy, which in the end leads to a coherent research. Saunders et al. (2019) use three different types of assumptions, ontology, epistemology, and axiology, when differentiating between the different research philosophies. Ontological assumptions mainly focuses on the reality's nature, and they account for the way you are seeing the world, which shapes the research. On the other hand, epistemological assumptions focuses on knowledge. More specifically, the legitimacy, acceptability, and validity of knowledge. Lastly, axiological assumptions mainly constitute the influence of ethics and values. This means that a researcher's values are SHAPing the way a research is developed. Saunders et al. (2019) present the following five research philosophies, positivism, critical realism, interpretivism, postmodernism, and pragmatism, in the business and management field.

In the context of this thesis, the authors assume a reality where the nature of an entity or event is not externally given. Especially the definition of hate speech is the result of a permanent debate in which meaning is constantly assigned and challenged. Additionally, refugees are also considered to be discursively constructed entities. This would normally lend itself to the domain of social constructionism. In contrary, critical realism focuses on the reality's underlying structures that shape observable events that can be seen and experienced. Unobservable structures, therefore, cause observable events resulting in the view that the social world can be understood if the structures that generate events are understood. Critical realism, thus, lends itself to the analysis of these events based on historical data allowing for a range of methods and data types, which are considered reasonable (Saunders et al. 2019, Sekaran and Bougie 2016). Accordingly, critical realism is the most fitting philosophical view for this thesis.

4.1.2 Research Approach

According to Saunders et al. (2019), research can be developed using one of the following three approaches, deduction, induction, or abduction. By relying on a deductive approach, the focus of the research is to develop a theory based on previous academic literature, where a research strategy is designed to perform multiple tests of the theory. Contrarily, an inductive approach is taken if the research develops a theory based on data that is collected to analyze a specific phenomenon. The abductive approach includes a mix of both deduction and induction, where it is possible to continuously move between these two. More specifically, this means that data is collected to both explain patterns and analyze a specific phenomenon, which is used to either develop a new theory or modify an existing one. Furthermore, this is tested using supplementary data to test the validity of the findings.

This thesis mainly relies on an abductive approach, as online opinionated text instances from multiple sources are collected and combined to explore the phenomenon of hate speech surrounding refugees using deep learning. Existing research is used as a starting point and the results of this specific research are used to expand the established research field. The validity of the results is further tested using additional data sources.

4.1.3 Methodological Research Choice

This next stage focuses on whether quantitative, qualitative, or mixed methods are used as a design of the research. Quantitative methods include numerical data, whereas qualitative methods deals with non-numerical data. These top-level categories can be further divided into both mono- and multi-methods. Using a mono-method means that only one data collection technique and analytical procedure are applied in the research, and the purpose of multi-method is to use multiple of these within either the quantitative or qualitative class separately (Saunders et al., 2019). Besides these two types of methods, there is also a mixed methods class that allows for an integration of both quantitative and qualitative techniques and procedures.

This thesis uses a mixed research method as both quantitative and qualitative techniques and procedures are applied. More specifically, the qualitative part originates from the col-

lected textual data, and a quantification of the textual data is needed to perform a statistical analysis, which leads to the use of quantitative procedures as well.

4.1.4 Research Strategy

A research strategy is one of the key elements needed in order to successfully conduct research. Such a strategy outlines the specific actions that are required to answer a defined research question. More specifically, it acts as a methodological connection between the philosophy and methods (Saunders et al., 2019). To clearly define a strategy, the research question and associated goals are taken into consideration together with the philosophy, approach, and methods. Besides that, the availability of existing knowledge and resources will also guide the choice of strategy. The following research strategies are discussed by Saunders et al. (2019), experiment, survey, archival and documentary research, case study, ethnography, action research, grounded theory, and narrative inquiry. This specific thesis primarily uses an experimental strategy as the goal is to construct a well-performing hate speech detection model in the context of refugee-related discussions.

4.1.5 Research Time Horizon

Another factor to consider when developing research is the time horizon as it influences the design of the study. Saunders et al. (2019) distinguishes between two types of time horizons, cross-sectional and longitudinal. A research is considered cross-sectional if it is working with a snapshot of a specific time, whereas a longitudinal research studies a phenomenon over a longer time period, which makes it possible to observe development and change over time (Saunders et al., 2019). The time horizon of this specific thesis is cross-sectional because change and development over time is not the core focus of the study. To obtain data, available datasets at the time of this study were collected and combined, which is further used for modelling. The data spans over multiple time periods, but in this research, it is treated as a single group, which makes it cross-sectional.

4.1.6 Techniques and Procedures

After having progressed through all the layers of the "Research Onion", the inner layer is reached. As a result, more practical choices have to be made, which includes the data collection techniques and analytical procedures needed to reach the objective of the research. Here, it is critical to establish an alignment between this stage and all the other stages

such that a successful research setup can be achieved (Saunders et al., 2019). More specific explanations of the applied data collection techniques and analytical procedures will be presented in the subsequent sections.

4.2 CRISP-DM

The methodology of this work primarily follows the Cross Industry Standard Process for Data Mining (CRISP-DM) framework proposed by Chapman et al. (2000), which presents an overview of a typical data mining project life cycle. It consists of a non-rigid sequence of six phases, where it is possible to move back and forth between each phase in the framework due to the iterative nature of a data mining project. The six phases are: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. These phases will be further explained in the subsequent sections.

The Business Understanding was performed in collaboration with UNHCR to understand both the needs and the problem that the organisation is trying to solve. As a result, based on these discussions, it was possible to align expectations. Furthermore, the literature review is also partly referred to as Business Understanding because it provides an overview of the different methods and technologies that have been used within similar works. This phase will not be further explained below like the other ones because it is introduced here and in earlier sections.

Figure 6: CRISP-DM Framework from Chapman et al. (2000)

The most frequent and important dependencies between the phases in the framework are represented by the arrows in Figure 6 (Chapman et al., 2000). The cyclical nature of a project is visualized by the outer circle. Since this framework is more suited towards the industry rather than the academic field, the Deployment phase is considered as out of scope for this thesis, and instead, it will be handled by UNHCR as the solution has to be integrated into their internal systems. However, this thesis recommend certain action points that can be implemented based on the results of the experiments.

4.3 DATA UNDERSTANDING

After having obtained a "Business Understanding", the next phase in CRISP-DM is "Data Understanding", which in this thesis consists of a data collection and description step. For all data sources, we explain the collection strategy and selection criteria together with providing a detailed description. Figure 7 shows how this steps fits into the overall process and methodology of this thesis.

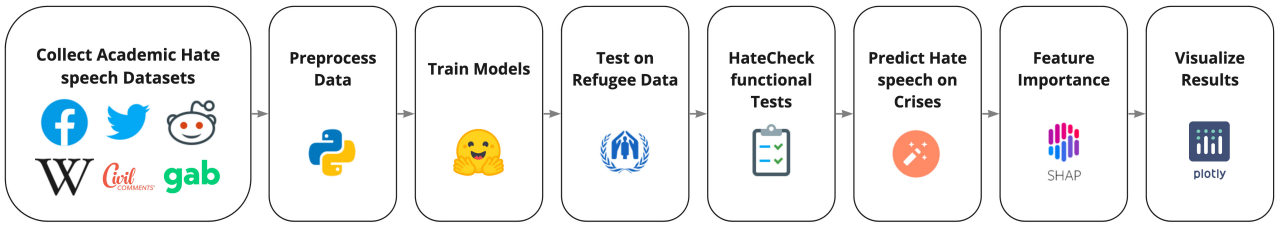

Figure 7: Methodology Overview

4.3.1 Data Collection

The data collection phase consists of multiple steps as different types of data are needed to answer the research and supportive questions. First, multiple datasets have been collected and combined into one large dataset that is used to train the deep learning models. More specifically, these datasets contain hate and offensive speech from multiple contexts such that more general hateful and offensive data can be used to train the models. A subset of that dataset is used to evaluate the models' performance on general hate and offensive speech. A thorough explanation of this is available in section 4.3.1.1. Besides that, a more specific type of data is needed to study how well hate and offensive speech can be measured in the context of refugees. This is introduced in section 4.3.1.2. Lastly, to analyze how hate

and offensive speech surrounding specific international refugee crises are changing over time, datasets for multiple crises have been obtained. This is explained in section 4.3.1.3.

4.3.1.1 General Hate Speech Datasets

Detecting and classifying the many forms of online abuse is a complex and nuanced task, which has proven remarkably difficult, both by humans and machines (Davidson et al. 2017; Waseem and Hovy 2016; Poletto et al. 2021). To create more robust, generalizable, and nuanced classification systems, the lack of clearly annotated, large, and detailed training datasets must be overcome. However, creating such datasets is time-consuming, complicated, and expensive (Vidgen et al., 2021).

By investigating the academic literature, it was discovered that the number of hate speech related datasets focusing on refugees was limited. Because of this, multiple academic datasets are combined into one large dataset, while ensuring that the same understanding of what constitutes as hate and offensive speech is maintained. The motivation behind this approach is to test whether it is possible to take advantage of hate speech datasets that have been used in different contexts than for refugees to detect hate speech targeted at refugees. In this section, the strategy and criteria for selecting and combining different datasets into one large dataset then used for training deep learning models are outlined.

SELECTION CRITERIA To control the quality of the data sources, a variety of criteria such as origin, actuality, data collection strategy, labels, and annotation process were considered. Regarding the origin of the text data, the type of platform where the text was posted was taken into consideration such that it was possible to obtain the desired type of text data. Besides that, the actuality of the text was also assessed to avoid obsolete data by considering the source and annotation date. Furthermore, the methods that have been used by other researchers to obtain data were also considered to filter data sources based on the data collection strategies that were applied. More specifically, the focus was to figure out whether the data was queried from for example Twitter using a seed-based strategy or collected from specific forums that have been deemed hateful.

Regarding the labels, both the type, quality, and number of labels were considered in the assessment of the datasets' quality. The goal of the data collection phase was to obtain multiple hateful and offensive datasets, which is why this criterion had a large impact on the assessment of each data source. The definition of each label in the datasets was crucial because it had to be aligned with the definitions that this thesis relies on. If definitions are not aligned, it can potentially confuse the models as there will not be a clear definition of hate and offensive speech. Moreover, the granularity of the labels was also considered.

Furthermore, the annotation process that was used to obtain labels for each of the data sources was weighted high when assessing the quality because it indicates the reliability of the labels. To assess this, multiple aspects were considered such as annotation guidelines, inter-annotator agreement scores, the number of annotators, and the types of annotators. The quality of the annotation guidelines is crucial because they provide information regarding the definitions of each label and how annotators should decide in specific situations. If the quality of these guidelines is poor, it is often reflected in the dataset as the annotators are not educated properly. Besides that, the type of annotators was taken into consideration because the quality of the labels is often related to whether the annotators are considered domain experts, random people or the researchers themselves. The most preferable are datasets annotated by domain experts, but these are sparsely available. As a result, the combined data includes samples that are annotated by all three types of annotators. The inter-annotator agreement was also considered in relation to the number of annotators such that it was possible to obtain datasets that had an acceptable agreement score. The acceptability of the score was assessed on a case-by-case level.

By relying on all of these different criteria, 12 distinct datasets were collected and combined into one large dataset. In section 4.3.2.1, these data sources will be presented. An id for each dataset has been constructed such that it is possible to refer clearly to each of them. This mapping is available in Table 4 together with details about each data source.

4.3.1.2 Data Surrounding International Refugees

To answer the research question and detect how well deep learning can be used to measure hate and offensive speech surrounding refugees, relevant data has been provided by UN-

HCR. The data has been collected using a commercial system that takes multiple key words as input, which are used to query the Twitter API for the desired type of data. More specifically, an internal system was used to construct the Twitter query using sudo-keywords that were provided by the employees.

4.3.1.3 Data Surrounding Specific International Refugee Crises

To both analyze and monitor the amount of hate, offensive, and normal speech on a refugee crisis level, additional data was obtained from Wolters and Olšavský (2021), who analyzed central events surrounding refugees. The goal of the research was to investigate how refugees are framed on social media. The data was collected using the Twitter API and contains tweets from the period January 2020 to April 2021. A seed-based approach was used to extract the tweets from Twitter by relying on the following query:

> *refugee OR refugees OR migrant OR migrants OR immigrant OR immigrants OR (asylum AND (seeker OR seekers)) OR ((displaced OR stateless) AND (people OR person OR persons))*

Within the large portion of collected data, Wolters and Olšavský (2021) identified multiple events connected to international refugee crises. As a result, a dataset was created for each of these events. In this thesis, datasets regarding the Afghanistan, United Kingdom channel crossings, Greece-Turkey, Rohingya and Tigray crises have been obtained.

4.3.2 Data Description

As the primary focus of section 4.3.1 is to explain how the different types of data have been collected, this section focuses more on describing the datasets.

4.3.2.1 General Hate Speech Datasets

In Table 4, an overview of the 12 different datasets that were obtained from a variety of sources is available. Each of these datasets are described more detailed below.

Dataset ID	Relevant Classes	Collection Strategy	Instances	Domain	Authors
cad	identity-, affiliation-, personal-directed abuse, non-hateful slurs, counter speech	Offensive Reddit communities	27,494	Reddit	Vidgen et al. (2021)
civil	toxicity, severe toxicity, identity attack, insult, obscene, threat	Annotated dataset	1,999,514	Civil Comments	Borkan et al. (2019)
davidson	hate speech, offensive	Beginning with the hatebase lexicon	24,783	Twitter	Davidson et al. (2017)
dynhs	hate speech, type of hate speech, target	Synthetic data from trained annotators	41,144	Adversarial	Vidgen et al. (2020)
ghc	human degradation/dignity, vulgar/offensive, call for violence	Random Sampling	27,546	Gab	Kennedy et al. (2022)
hasoc	hate speech, profane, offensive	Seed-based	7,005	Facebook, Twitter	Mandl et al. (2019)
hatemoji	hate speech	Synthetic data from trained annotators	5,912	Adversarial	Kirk et al. (2021)
hatexplain	hate speech, offensive	Lexicon-based	20,098	Gab, Twitter	Mathew et al. (2020)
hateval	hate speech, aggressive	Monitoring accounts, seed-based & user history	10,000	Twitter	Basile et al. (2019)
ousid	hate speech, offensive, disrespectful	Seed-based	5,647	Twitter	Ousidhoum et al. (2019)
slur	derogatory	Filtering for slurs	39,811	Reddit	Kurrek et al. (2020)
wikipedia	toxic, severe toxic, obscene, threat, insult, identity hate	Annotated dataset	223,549	Wikipedia	Wulczyn et al. (2016)

Table 4: Overview of datasets used for the combined dataset

CONTEXTUAL ABUSE DATASET (CAD) The Contextual Abuse dataset is based primarily on English entries from the social media site Reddit. Its main focus is to address the limitation of previous datasets. This is done by providing five distinct primary categories, which each have multiple secondary categories. The primary categories are further divided into an abusive category, which contains identity-, affiliation- and person-directed abuses. Besides that, they are also divided into a non-abusive category, which includes non-hateful slurs and counter speech. The secondary categories are identical for the primary categories identity- and affiliation-directed abuse. They make a distinction between derogation, animosity, threatening language, and dehumanization. Labeled posts and comments can have multiple primary and secondary categories. Only the category of identity-directed abuse was considered for this thesis. (Vidgen et al., 2021).

Type of Speech	Text Example	n
Derogation	"People that are born without legs are not people. We need a separate category for them."	982
Threatening language	"If they are not afraid of us clearly we need to give them something to fear "	28
Dehumanization	"White arent people, they're a disease yakub unleashed onto the land"	27

Table 5: Example texts from the CAD dataset for each of the relevant classes. Only the category of identity-directed abuse was considered.

As the widely used practice of keyword sampling can introduce topic and author biases (Wiegand et al., 2019), the sampling strategy for the CAD dataset is community-based. In total, 117 subreddits where selected, which have a higher likelihood of containing different levels and divergent use of abuse, leading to a more realistic dataset. From the dataset, the identity-directed abuse is only relevant for this thesis. Furthermore, the dataset provides high quality annotations as the data is annotated by an expert-driven group-adjudication process, where all entries were independently annotated by at least two annotators, who underwent four weeks of training and were either native English speakers or fluent. All disagreements were surfaced and discussed with an expert with the goal to improve the annotators understanding. The inter-annotator agreement for the primary categories using Fleiss' Kappa (Fleiss, 1971) was $\kappa = 0.58$. In Table 5, examples for each of the relevant classes that are considered in this thesis are available.

CIVIL COMMENTS DATASET The Civil Comments dataset provides seven primary labels focused on toxicity in online conversations coming from an archive of the Civil Comments platform, a commenting plugin for independent news sites. The comments were created on approximately 50 English-language news sites across the world between 2015-2017. This dataset was annotated by and released for the Jigsaw Unintended Bias in Toxicity Classification Kaggle challenge. The dataset is unique in its size with $\sim$ 2M annotated comments. The goal of the dataset was to identify toxicity in online conversations by using the following definition "toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion"; (Borkan et al., 2019). Providing labels such as severe toxicity, insult, obscene, threat, and identity attack makes the dataset highly useful for researching abusive language phenomena. The annotation process was crowdsourced, and each comment was shown to around 10 annotators, who were screened for rating accuracy (Borkan et al., 2019). In Table 6, examples for each of the relevant classes that are considered in this thesis are available.

DAVIDSON Davidson et al. (2017) utilized a hate speech lexicon from *Hatebase*, which consisted of phrases and words that have been provided by Internet users. The terms from this lexicon were used to query the Twitter API for tweets. Additionally, the timeline of the

Type of Speech	Text Example	n
Identity Attack	"not all Muslims are terrorists, but all terrorists are muslim."	8,402
Insult	"Yes. It's a small group, but they are EXTREMELY whiny..."	88,575
Obscene	"I can't wait to throw that bum Seaton out."	8,482
Threat	"My motto? "Support mental health or I'll kill you and your family.""	3,084
Neither	"Hm-mm..aren't some people claiming the Navy presence has been a real economic drag on the area?"	1,898,502

Table 6: Example texts from the Civil dataset for each of the relevant classes.

users that wrote these tweets were also collected, which gave 85.4 million tweets. Out of these, 25k was randomly selected and annotated by workers from CrowdFlower. Based on a description and definition provided by the authors, the workers had to determine whether a tweet was considered as offensive, hate speech or neither. The tweets were annotated by at least three workers, and an inter-annotator agreement score of $\kappa = 0.92$ was achieved. This score was provided by the annotation platform. Out of the annotated tweets, around 5% were labelled as hate speech, whereas most of the tweets were labelled as offensive. In Table 7, examples for each of the relevant classes are available.

Type of Speech	Text Example	n
Hate	"@Huntermoore happy cuz I woke up white and not a [n-slur]!"	1,430
Offensive	"!!!!!!!!! [...] she look like a tranny "	19,190
Neither	"Yankees take all three games in MoTown, I'm callin' it!!! MLB "	4,163

Table 7: Example texts from the Davidson dataset for each of the relevant classes.

DYNHS Vidgen and Derczynski (2020) used a novel human-in-the-loop process to dynamically generate high amounts of synthetic hate speech data with fine-grained labels. The data is generated and labeled by trained annotators in four rounds of creation. Besides that, the data also includes challenging perturbations. The dataset is especially interesting as 54% of the examples are hate speech, a much higher percentage than most datasets. In the dynamic creation, annotators are not labelling data. Instead, they are producing high-quality adversarial samples, which are challenging for baseline models. This process is repeated iteratively over four rounds, and offers a much more targeted way of collecting data. Twenty annotators in total, overseen by two experts, were extensively trained to

validate synthetic examples and perpetrations that fooled the model in each round. Each example was validated by three to five annotators. To measure the agreement of the annotators, Krippendorf's alpha was used, and the score varies in the different rounds with values between $\alpha = 0.45$ and $\alpha = 0.82$. In Table 8, examples from the relevant classes that are considered in this thesis are available.

Type of Speech	Text Example	n
Hate	"gay people are a burden to our society"	22,175
Not Hate	"it is false that all gay people are cunts"	18,969

Table 8: Example texts from the Dynhs dataset for the relevant classes.

GAB HATE CORPUS (GHC) The Gab Hate Corpus is collected by Kennedy et al. (2022), and contains 28k randomly sampled posts from Gab in the time period January 2018 to October 2018, which are annotated into multiple hate-based labels. More specifically, the posts were assigned to the following three top-level categories; Human Degradation, Calls for Violence, and Vulgar/Offensive. Besides that, the dataset also contains labels regarding the group that is being targeted in the posts. The annotation was performed by undergraduate research assistants, who were provided with annotation guidelines that contained examples, definitions, and a discussion of categories for guidance. Besides studying this manual, the annotators further had to pass a test, where they had to annotate a subset of posts provided by the authors. For the final dataset, each post was annotated by at least three of the annotators and a $\kappa = 0.27$ score was achieved for the top-level categories. In Table 9, examples for each of the relevant classes that are considered in this thesis are available.

Type of Speech	Text Example	n
Call For Violence	"Slaughter all who have a head slope less than 70 degrees!"	155
Human Dignity	"Many men in the west are very weak men. And you are one of them Frank."	2,348
Vulgar/Offensive	"I do know that you're retarted enough not to be able to write a coherent sentence."	1,748
Neither	"Just saw an Oreo cookie commercial. Drool."	24,169

Table 9: Example texts from the GHC dataset for the relevant classes.

HASOC For the HASOC track at FIRE 2019, a dataset that contains text data from both Twitter and Facebook was provided to the participants in German, English, and Hindi (Mandl et al., 2019). The English part of the dataset, which consists of 7,005 text examples

that had been annotated in three different categories by several juniors, was extracted to be part of this project's combined dataset. The tweets were obtained by querying the Twitter API for different keywords and hashtags that were considered offensive and hateful. Additionally, the latest posts from the authors that had posted offensive and hateful content were collected too. The first label category consists of two types of labels, hate and offensive versus non-hate and offensive. The second category contains the following three labels; hate speech, offensive, and profane. The last category focuses more on whether the hateful and offensive text is targeted or not. Based on the annotation process, the three sub-tasks achieved an inter-annotator agreement score of 0.78, 0.69, and 0.75 respectively. In Table 10, examples for each of the relevant classes that are considered in this thesis are available.

Type of Speech	Text Example	n
Hate	"Omar is a disgrace to Islam."	1,267
Offensive	"A good morning and a fucktrump to you, Mr. Fuckus!"	522
Neither	"I think majority are just giving him the benefit of doubt let's see what he does and if we are out by 31oct"	5,216

Table 10: Example texts from the Hasoc dataset for the relevant classes.

HATEMOJI Building on the human-in-the loop approach from Vidgen and Derczynski (2020), and on the idea of generating high quality hate speech data through adversarial examples, Kirk et al. (2021) provide a dataset consisting of 5,912 examples that addresses the weakness of models to detect hate expressed with emojis. Seven different functionalities are contained in the dataset. For example, one functionality swaps parts of a sentence with an emoji like a verb, an identity or a descriptor for dehumanizing language. All the swaps were combined so that in some samples, the text only consists of emojis. Moreover, in some cases non-hateful statements were appended with negative emojis, and in other cases clearly hateful statements were appended with positive emojis to make them hateful. Twitter analysis was used to identify these emoji swaps. In total, three rounds of generation and synthetic perturbations were carried out to create the dataset. Furthermore, the dataset is accompanied by HatemojiCheck, a test set of further 3,930 short-form statements that makes it possible to evaluate the performance based on the seven aforementioned functionalities, extending the work of HateCheck provided by Röttger et al. (2020). In Table 11, examples for each of the relevant classes that are considered in this thesis are available.

Type of Speech	Text Example	n
Hate	"I am working against 🏧 cause they are fucking ghastly people"	2,965
Not Hate	All these black guys round here are genuinely 😇, great vibes	2,947

Table 11: Example texts from the Hatemoji dataset for the relevant classes.

HATEXPLAIN The benchmark dataset for hate speech detection introduced by Mathew et al. (2020) is also collected to be part of the combined dataset. It consists of 9,005 tweets from Twitter and 11,093 posts from Gab. The tweets were collected in the period January 2019 to June 2020 using multiple lexicons from previous research. Additionally, the Gab dataset was obtained from Mathew et al. (2019). The labels are distributed into three categories. The first category indicates whether a text is hateful, offensive or neutral, whereas the second one accounts for the communities that are targeted in the text. The last label category contains the words and phrases that led the annotators to decide that a text was either hateful or offensive. Furthermore, each text in the dataset was annotated by 3 workers from Amazon Mechanical Turk, where they were provided with instructions, definitions, and examples to support their annotations. To ensure high quality, the qualification requirement from Amazon Mechanical Turk was used to select reliable workers. Besides that, a pilot task was used to further increase the quality of the annotators, where 621 workers were tested on a small subset of the data. Out of these, 253 possessed the skills to perform this annotation task. An inter-annotator agreement score of $\alpha = 0.46$ was achieved on the final dataset. More specifically, the score was based on Krippendorff's α (Krippendorff, 2011). In Table 12, examples for each of the relevant classes that are considered in this thesis are available.

Type of Speech	Text Example	n
Hate	"why are we having all these people from shithole countries coming here"	6,234
Offensive	"suck out you [f-slur]"	5,761
Neither	"smiling cuz i am young black rich and i am handsome"	8,103

Table 12: Example texts from the HateXplain dataset for the relevant classes.

HATEVAL The dataset that was used in the shared-task, SemEval 2019 Task 5, provided by Basile et al. (2019) was also extracted and used in this thesis. The shared-task primarily focused on detecting hate speech in English and Spanish tweets targeted at women and im-

migrants. Due to this thesis' topic delimitation, it was only the English text data that was extracted. The majority of the dataset was collected in the period July 2018 to September 2018 using multiple data collection strategies. More specifically, potential victim accounts were monitored, the history of tweets from identified haters were obtained, and lastly, a seed-based strategy using keywords, stems, and hashtags was used to query Twitter. In total, 19,600 tweets were collected, where 10,509 of them is about women, and 9,091 is about immigrants. Out of the total number of tweets, 10,000 of them were written in English and used as a training set. To obtain labels for hate speech, the target (group/individual), and aggressiveness in hateful tweets, contributors from Figure Eight, a crowd-sourcing platform, were used. The annotators were provided with annotation guidelines that contained information about definitions together with examples. Furthermore, at least three people annotated each tweet, and an average confidence score[1] for the English tweets of 0.83 (hate speech), 0.70 (target), and 0.73 (aggressiveness) were obtained. Additionally, two expert annotators were used, and to obtain the final labels, majority voting was used. In Table 13, examples for each of the relevant classes that are considered in this thesis are presented.

Type of Speech	Text Example	n
Hate	"This immigrant should be hung or shot! Period! Animal"	4,210
Not Hate	"Merkel under pressure to defuse dispute over migration"	5,790

Table 13: Example texts from the Hateval dataset for the relevant classes.

OUSID A multi-lingual dataset that covers multiple aspects of hate speech was created by Ousidhoum et al. (2019). The English part of the dataset is extracted, and used in this work. The text dataset was collected from Twitter using a seed-based strategy by relying on keywords and phrases. The total English dataset consists of 5,647 tweets. To increase the reliability of the annotators, guidelines and aligned label sets were handed out to potential annotators, and only the ones with an accuracy higher than 90% were chosen for this task. Regarding the labels, there are five different categories in the data; directness, hostility, target, group, and annotator sentiment. Each of these categories contain a number of more fine-grained labels. In this thesis, it is only the hostility category that is used. More specifically, it accounts for a multi-label task that describes whether a tweet is abusive,

[1] A score that combines the contributors reliability together with the inter-annotator agreement (Basile et al., 2019)

fearful, offensive, hateful, disrespectful and/or normal. All the tweets were annotated by five people and majority voting was used to obtain the final label. Due to the fined-grained nature of all the different types of labels, an inter-annotator agreement score of $\alpha = 0.15$ was achieved. More specifically, the score was based on Krippendorff's alpha (Krippendorff, 2011). In Table 14, examples for each of the relevant classes that are considered in this thesis are available.

Type of Speech	Text Example	n
Hateful	"@user swear .. white people fucking retarded ..."	1,278
Disrespectful	"i'm forced deal retarded people everyday"	682
Offensive	"[...] i'm making buttered noodles u retard"	3,920
Neither	"@user visited taiwan once!i hope go back country someday"	778

Table 14: Example texts from the Ousid dataset for the relevant classes.

SLUR Kurrek et al. (2020) filtered for slurs in the Pushshift Reddit corpus that contained data from the period October 2007 to September 2019 provided by Baumgartner et al. (2020). After having performed author, comment, and community level filtering, a corpus of 40,000 comments was obtained. The dataset consists of four major labels; derogatory usage, appropriative usage, non-derogatory and non-appropriative asage, and homonyms, which are further split into 12 sub-labels. However, these sub-labels are not used in this thesis. All the labels were annotated by 20 people coming from the university environment, who was trained for the task in a workshop running for two days. Each comment in the dataset was annotated by two annotators, and the ones where the annotators disagreed were handled by the authors. Moreover, a few comments were identified as noise, which were removed, and as a result, the final dataset consists of 39,811 comments, where a Cohen's Kappa (Cohen, 1960) of $\kappa = 0.60$ was achieved. This was deemed sufficient compared to the related literature. In Table 15, examples for each of the relevant classes that are considered in this thesis are available.

WIKIPEDIA TALK CORPUS The Wikipedia Talk comments dataset contains 223k labeled discussion comments from an archive of English Wikipedia talk page. Like the Civil Comments dataset, the annotation was done by the Alphabet company, Jigsaw, and first released

Type of Speech	Text Example	n
Derogatory	"Christina should sit back like a good tranny and let the real women have their day."	20,530
Not Derogatory	"[N-slur] isn't a race either. It's a derogatory term that refers to blacks, the same way guido is"	19,281

Table 15: Example texts from the Slur dataset for the relevant classes.

in the Jigsaw Multilingual Toxic Comment Classification Challenge. The labels are similar to the ones from the Civil Comments dataset, with multiple languages being available. For the annotation process, approximately 10 annotators were used to classify each comment via CrowdFlower focusing on personal attacks, aggression and toxicity as per goal of the Wikipedia Detox project (Wulczyn et al., 2016). Using Krippendorf's alpha, Wulczyn et al. (2017) measure an inter-annotator agreement score of $\alpha = 0.45$. In Table 16, examples for each of the relevant classes that are considered in this thesis are presented.

Type of Speech	Text Example	n
Identity Hate	"All the admins on here are fucking [n-slur]. I hate [n-slur], they are all retards. This proves my point that [n-slur] want to take over the world. So fuck all you [n-slur]. [n-slur] out of the United States. Whites rule. [...] "	2,007
Insult	"::No, he is an arrogant, self serving, immature idiot. Get it right."	11,304
Obscene	"HE IS A GHAY ASS FUCKER@@!! "	12,140
Threat	"I am going to rip off your tiny balls then decapitate you."	689
Neither	"She was born between December 22 and december 31 in 1982. Just so you know. I'll put December 1982. "	208,498

Table 16: Example texts from the Wikipedia dataset for the relevant classes.

4.3.2.2 *Data Surrounding International Refugees*

The data that was obtained from UNHCR consists of 98 tweets that solely contain hate speech against refugees, which has been flagged in their systems. Some of the text instances were originally written in a different language than English, and as a result, it has been translated into English using Google Translate by UNHCR to match the language constraint specified for this project. Regarding the annotation of the text, it has been annotated by a single employee from UNHCR. Besides the tweets, multiple meta-data was provided as well, but since this is out of the scope for the thesis, it is left out. The primary idea with this specific dataset is to use it to test deep learning models' ability to classify hate and

offensive speech within the domain of refugees. Multiple example tweets from the dataset can be found in Table 17.

Tweets Surrounding Refugees
"#RefugeesWelcome is a terrorist organisation that is helping BORIS's government with their Great Replacement Policy to replace all British people with the scum of the earth."
"@USER @USER Most are not refugees, they are bogus refugees, economic migrants looking for a better life, some are criminals and some are terrorists."
"@USER That's what happens when you transport thousands of unvetted refugees. Basically bringing extremism and terorists to our own doorstep."

Table 17: Example tweets from the UNHCR refugees test dataset. All samples are hate speech.

4.3.2.3 Data Surrounding Specific International Refugee Crises

Details regarding the datasets are provided in Table 18. Besides the textual part of the dataset, it also consists of multiple types of meta-data, which are discarded due to the irrelevancy according to this thesis. The data is a bit different than the data that is presented above, because it does not contain any labels. The purpose of using these datasets is to analyze the development of hate, offensive, and normal speech within these crises over time. This allows us to potentially identify different patterns associated with the online discussion surrounding refugees.

By creating a model that is able to classify text as being either hateful or offensive makes it possible to monitor the development of the public opinion of refugees. From the perspective of UNHCR, this can support their communication team because if the amount of hate speech rises, the team can take preventative controls in use to steer the online discussion. The datasets that are used here, only constitute a subset of the overall Twitter discussion surrounding these crises, which is why these are mainly used for proof-of-concept purposes. To avoid redundancy, only the Afghanistan and Greece-Turkey crises will be analyzed in the results of this thesis. Therefore, it is only these two crises that are introduced below. The graphics and tables presented in the results are available for the other crises in section A.3. In the summer 2021, the humanitarian situation in Afghanistan deteriorated considerably as a result of the United States withdrawing its military forces from the country (UNHCR, 2022b). This event forced additional 175 thousand people to become refugees, resulting in 2.4 million Afghan refugees in total. The vast majority of these refugees departed

Crisis	Start Date	End Date	n tweets
Afghanistan	26-07-2021	31-08-2021	283,643
Channel Crossing	19-07-2020	29-08-2020	173,758
Greece-Turkey	11-02-2020	23-03-2020	137,462
Rohingya	01-03-2021	30-04-2021	29,432
Tigray	15-01-2021	30-04-2021	42,853

Table 18: Overview of the datasets surrounding international refugee crises obtained from Wolters and Olšavský (2021).

Afghanistan throughout the years, beginning in 1979 and are mostly registered in Iran and Pakistan, which are Afghanistan's neighbors. UNHCR also issued a non-return advice for Afghanistan in August 2021, calling for a halt of forcible repatriation of Afghan citizens, including asylum seekers whose claims have been denied. The data used for the analysis is centered around the 15th of August 2021, the date of the fall of Kabul (UNHCR, 2022a).

The refugee crisis associated with Greece and Turkey is a result of the dispute between these two countries. An armed conflict was taking place in Syria in 2020, therefore, it was expected that thousands of people would flee from their homes in Syria. As a result, the Turkish President, Recep Tayyip Erdoğan, announced that refugees were allowed to enter Europe using the border between Greece and Turkey (Stevis-Gridneff and Gall, 2020). The reaction to this decision from Greece was to implement different kinds of measures to avoid the influx of refugees. As a result, thousands of refugees were stuck at the border as it was not possible to continue their journey into Europe. To analyze the development of the volume of hate, offensive, and normal speech over time for this specific refugee crisis, Twitter data has been obtained from the period 11th of February 2020 to 23th of March 2020, where the announcement from the Turkish President was published.

4.4 DATA PREPARATION

After having obtained a more thorough understanding of the datasets that are used in this thesis, the next phase in CRISP-DM is "Data Preparation". Multiple preparation steps have to be implemented to achieve, from a modelling perspective, the desirable state of the collected data. Since these datasets originate from different sources, it requires some specific preprocessing of the data to align the format among all the datasets. Moreover,

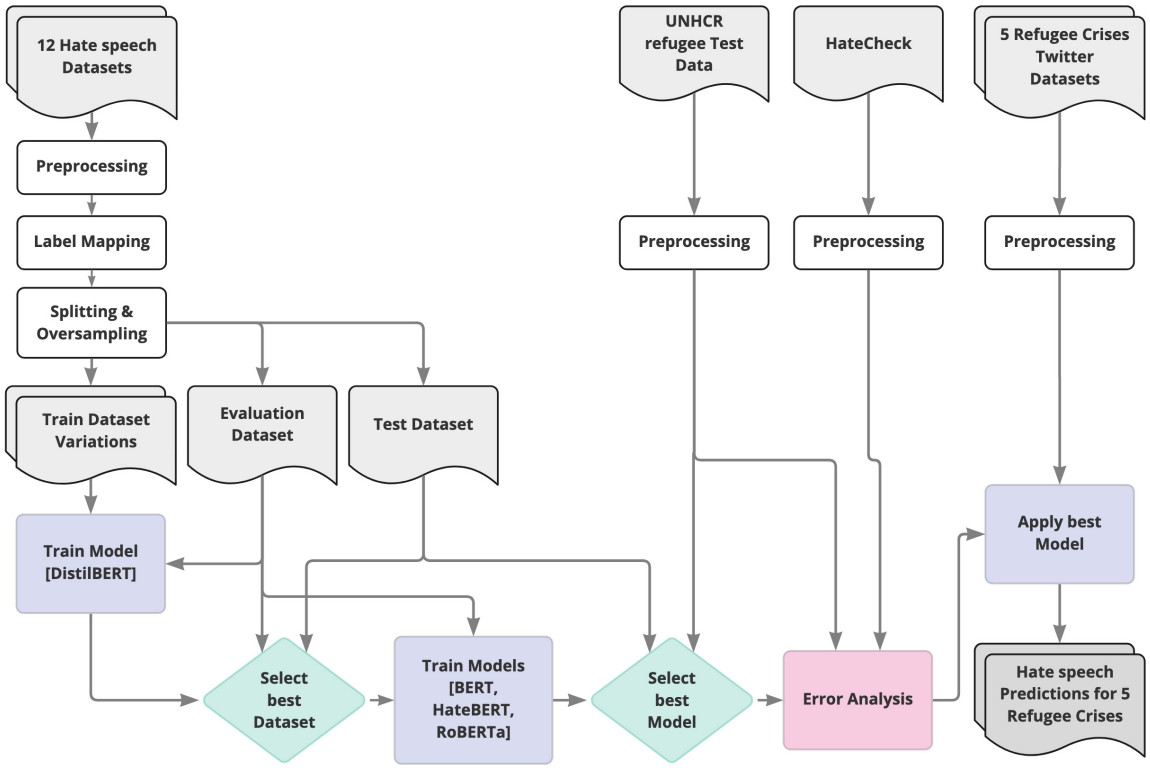

Figure 8: Data processing and model selection flow.

the techniques that have been used to handle the data and labels also vary between each dataset, which is why a standardized format is needed. This will be further explained in the following sections. Figure 8 shows how the various data preparation steps fit into the overall data and modelling flow.

4.4.1 Data Preprocessing

To obtain a standardized format of all the datasets, the preprocessing techniques that have been applied by Caselli et al. (2021) are used together with some additional ones. By relying on similar techniques, it provides some sort of consistency as this thesis further fine-tunes the HateBERT model proposed by Caselli et al. (2021), which will be introduced later. The most prominent techniques that have been used to preprocess the textual data are visualized in Table 19.

First, all the text is transformed into a lowercase representation to eliminate differences in the use of capitalization. This means that a word, which initial letter is capitalized is now identical to the same word which only consists of non-capitalized letters. Since most of the data originates from social media platforms, multiple user mentions were found in the

Cleaning Steps	Text
Raw Text	This immigrant should be hung or shot! Period! #Animal. @example_user 😂 https://t.co/wFcGoLCqJ5
Lowercase	this immigrant should be hung or shot! period! #animal. @example_user 😂 https://t.co/wfcgolcqj5
User Mentions	This immigrant should be hung or shot! Period! #Animal. @USER 😂 https://t.co/wFcGoLCqJ5
URLs	This immigrant should be hung or shot! Period! #Animal. @example_user 😂 URL
Hashtags	This immigrant should be hung or shot! Period! Animal. @example_user 😂 https://t.co/wFcGoLCqJ5
Emojis	This immigrant should be hung or shot! Period! #Animal. @example_user :joy: https://t.co/wFcGoLCqJ5
Preprocessed Text	this immigrant should be hung or shot! period! animal. @USER :joy: URL

Table 19: The pre-processing steps for textual data.

text data. To handle this, every time a user is mentioned in the text it is converted into a @USER token. Besides that, many posts contain an URL to either former tweets or websites. These are often in the form of shortlinks, and thus, are not useful for the classification task at hand, see example in Table 19. As a result, these links are transformed into an URL token. Furthermore, hashtag characters are removed from the text data to free up words as regular text. By exploring the data, it also became evident that multiple white spaces were present. To handle this, these additional spaces were replaced with a single white space.

On social media platforms, there have been an increasing use of emojis when communicating. According to Alshenqeeti (2016), emojis can stand alone in a sentence without the presence of any words to represent the meaning. This suggests that emojis are influential features because they posses a lot of meaning. To take advantage of this, the Python library *emoji* is used to convert emojis into text. Each type of emoji is associated with a unique identifier in text format, which is inserted into the text instead of the emoji itself. This identifier describes the emoji, which can be observed in Table 19. Besides that, in the exploration of the dataset, multiple outliers were identified in the data from Wikipedia as the text samples contained far more words compared to the rest of the data. To handle this, a constraint was specified to filter out texts with more than 256 words.

4.4.2 Label Mapping

The transformations explained in this section are only applied to the dataset that combined 12 different general hate speech datasets. As outlined in Table 4, all datasets identified use different and conflicting annotation schemes. To be able to compare different labels across datasets, the label classes are standardized. More specifically, all the relevant labels are mapped into the following three classes; hate speech, offensive, and normal. However, the normal class is more of a neither class. To perform the standardization, the definition and annotation guideline for each label in each dataset were considered. After carefully checking if the definition of a label was equivalent with our definition of hate or offensive speech, it was assigned the new standardized class. Table 21 shows the standardization that was performed.

	Normal	**Offensive**	**Hate speech**	Total
Number of samples	263,406	135,190	74,023	472,619
Corpus size (word count)	11,152,326	4,859,976	1,967,928	17,980,230
Number of unique words	476,007	220,267	119,224	619,575
Mean number of words per sample	42	36	27	38

Table 20: Descriptive statistics of combined dataset after label mapping was performed.

All the labels that did not match with the definitions in this thesis were assigned to the normal class. Because the datasets `civil` and `wikipedia` contain the most amount of normal data with up to 3M rows of mostly neutral data, it was decided to limit the amount of normal posts from these two source datasets to 95k and 75k respectively to reduce the total size of the dataset. Furthermore, descriptive statistics for the combined dataset after the label mapping are available in Table 20. From the table, it can be observed that 74,023, 135,190, and 263,406 number of samples are available for the hate speech, offensive, and normal class respectively. Besides that, word length descriptive statistics for each source dataset are available in Table 35 in the Appendix.

Dataset ID	Original Class	Standardized Class
cad	derogation	hate speech
	threatening language	hate speech
	dehumanization	hate speech
Civil	identity attack	hate speech
	insult	offensive
	obscene	offensive
	threat	offensive
davidson	hate	hate speech
	offensive	offensive
dynhs	hate	hate speech
ghc	call for violence	hate speech
	human degradation/dignity	hate speech
	vulgar/offensive	offensive
hasoc	hate	hate speech
	offensive	offensive
hatemoji	hate	hate speech
hatexplain	hate	hate speech
	offensive	offensive
hateval	hate	hate speech
ousid	hateful	hate speech
	disrespectful	offensive
	offensive	offensive
slur	derogatory	hate speech
wikipedia	identity hate	hate speech
	insult	offensive
	obscene	offensive
	threat	offensive

Table 21: Label standardization. The table shows the original class from each dataset and its standardised class.

Figure 9 shows the distribution of labels for both the hate speech and offensive class for each of the 12 datasets that are combined. From the visualization, it can be observed that `slur` and `dynhs` provide the bulk of the hate speech data with over 20k entries each. Besides that, it can also be observed that only 7 datasets have contributed to the offensive class. However, most of the contribution comes from the `civil` dataset.

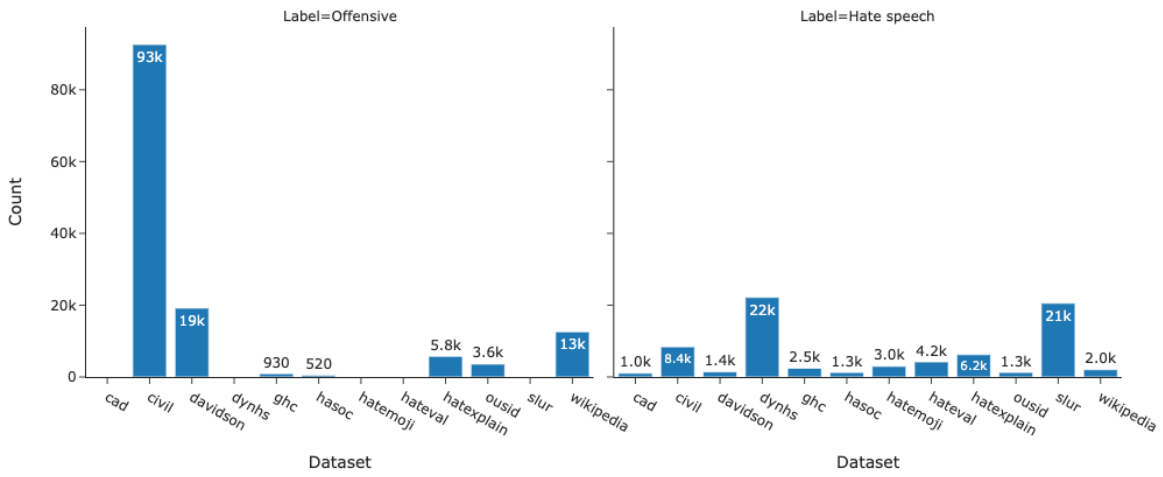

Figure 9: Contribution of source datasets towards the hate speech and offensive class.

4.4.3 Data Splitting

To be able to perform a valid evaluation of the performance of a deep learning model, the obtained dataset needs to be split into a train, validation, and test set. More specifically, the train set is used to fit the parameters of the classifier, whereas the validation set is used to tune parameters of a model and select the best model. On the other hand, the test set is only used in the end to asses the performance of a fully-specified model (Ripley, 1996). As the UNHCR refugee test dataset is not used for training, it does not have to be split. Instead, it is only the large dataset that consists of data from 12 different sources that needs to be split as it is used for training the deep learning models.

COMBINED DATA SET Usually, the data which a model is trained and tested on draws from the same source. In this thesis, this is the combined and standardized dataset consisting of the 12 different hate speech datasets as described in subsection 4.3.2. To create the split, the dataset was first shuffled and then a standard 80/10/10 split was applied. This means that 80% of the dataset is used for training, 10% for validation, and 10% for testing. Additionally, the split was stratified by labels and by datasets. This ensures that each label in combination with each dataset is preserved proportionally in all sets. It can be noted that the stratified split along two axes only tries to approximate the proportions, which given the total size of the dataset is sufficient for testing. This results in 378,095 examples for the training set, and each 47,262 for the validation and test set.

4.4.4 Dataset Variations

The combined training dataset consists of many different data sources that each have been validated through previous research. To find an optimal composition, multiple versions of the datasets were created and tested using a deep learning model. This is explained further in section 4.5.4. The challenges of imbalanced classification, meaning a dataset where the proportion of elements in each class significantly differ remain problematic in the field of NLP and Machine Learning in general. This makes it more difficult for the model to generalize on dissimilar data as the class of interest (e.g. offensive or hate speech) is significantly lower than other classes (Madabushi et al., 2020).

OVERSAMPLING The combined dataset only consists of 15.6% hate speech and 28.6% offensive data, therefore, it is considered imbalanced. Since the hate speech class is least represented in the dataset, a dataset variation was created, where the hate speech class was oversampled. Because of the varying quality of underlying datasets, and based on previous research of Vidgen et al. (2020), the oversampling was performed unequally over the training data. Hyperparameter optimization was used to find the best performing combination of multiples for certain datasets. As a result, the following datasets were added to the train dataset: {`dynhs`: 1x, `cad`: 4x, `hatemoji`: 5x, `ghc`: 3x, `hatexplain`: 2x}. The resulting distribution per dataset for the hate speech class can be observed in Figure 10. Overall, the training set of the oversampled dataset variation has 456,003 rows of training data with 132,964 examples being labeled as hate speech. This constitutes 29.2%.

4.5 MODELLING

This next phase in CRISP-DM introduces the different modelling frameworks that have been used to conduct the various experiments. The chosen models take advantage of the transformer-based architecture from Vaswani et al. (2017), and constitute variations of the $\text{BERT}_{\text{base}}$ model introduced by Devlin et al. (2018). The motivation for solely applying different variations of the BERT model is due to its proved robustness in a variety of natural language processing tasks, which recently have led to state-of-the-art performances. Besides introducing the modelling frameworks, the different variations of datasets, explained

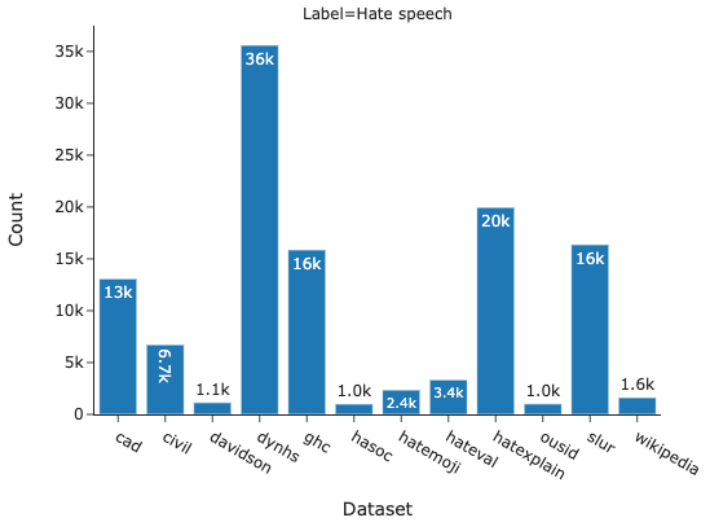

Figure 10: Distribution of source datasets for the hate speech class after oversampling

in section 4.4.4, will be tested such that the best performing dataset is used for further experiments.

4.5.1 Overview of Model Architecture

Using the BERT modelling architecture to classify hate and offensive speech requires multiple components. More specifically, an input sequence, tokenizer, language model, and a classification model is needed. The connection between these is visualized in Figure 11. This section introduces the different components on a high level as a more detailed description of how a BERT model works under the hood is provided in section 2.4.3.

The first step in the process is to feed the input sequences into a tokenizer to obtain a desired representation of the input data such that it fits a specific model. The tokenizers provided by Huggingface are model-specific, which means that each tokenizer is connected to a pre-trained language model. A tokenizer performs multiple transformations. First, WordPiece tokenization is performed, which takes the input sequence and splits it into word tokens or word pieces. Afterward, the language model requires sequences of equal length, therefore, the tokenizer also performs padding or truncation to ensure this. Furthermore, the tokenizer also adds the two special tokens, `[CLS]` and `[SEP]`, to the input sequence. The `[CLS]` token is used to make an aggregated representation of the whole sequence, which is later used to perform the classification task. The `[SEP]` token is primarily

used for the next sentence prediction task to separate two sentences (Devlin et al., 2018). Additionally, the attention masks are also created to ensure that the model is not going to attend the padded tokens. The last task of a tokenizer is to convert the tokens into ids to obtain a numerical representation of the textual data. Since the RoBERTa language model is applied in this thesis, it is important to state that it uses a different tokenizer that relies on Byte-Level Byte-Pair-Encoding (BPE) instead.

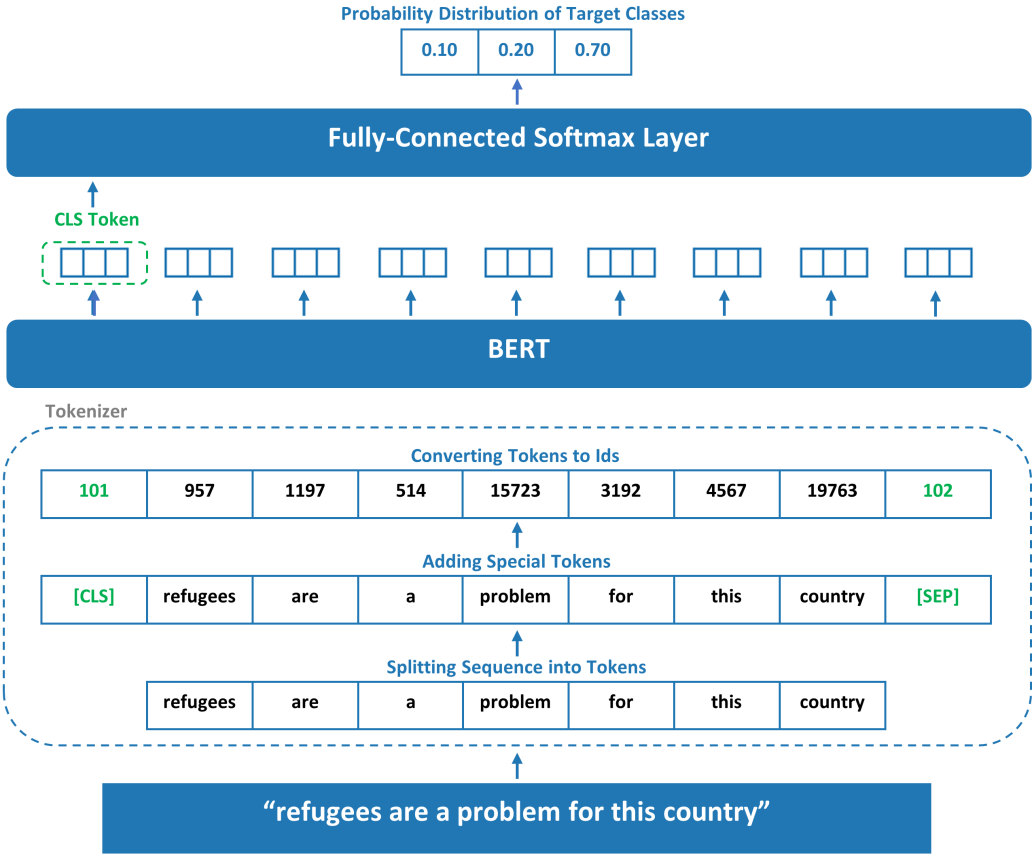

Figure 11: High-level model architecture of BERT inspired by Alammar (2019)

After the input sequences have been transformed using a tokenizer, it is added to the language model, which is treated as a black box here because the details of how it works have already been explained. The pre-trained language model is used to create a deep bidirectional representation of the input text, and outputs vectors of a specific hidden size that represents the text. Since the main focus of this thesis is a classification task, the final representation of the `[CLS]` token is more relevant. As mentioned above, the [CLS] token represents the whole sequence of an input, and to be able to perform a classification task,

the final representation of this token is fed into a fully-connected feed-forward neural network layer (Liu et al., 2019).

Based on the pre-trained language model, it is now possible to create a probability distribution over the different target classes by applying either a sigmoid or softmax activation function in the fully-connected layer. The type of activation function depends on whether the classification problem is considered binary (sigmoid) or multi-class (softmax). In this thesis, the classification problem is considered multi-class, therefore, it uses a softmax activation function. The output of the fully-connected layer is a vector with a dimension equal to the number of target classes, which consists of probabilities in the range $0-1$ that indicates which class is more likely than the other ones.

4.5.2 Specifications for Model Training

To be able to reliable predict hate and offensive speech, a deep learning model has to be thoroughly trained and optimized. This section introduces various details surrounding the construction and training of the chosen deep learning models, which includes information regarding the programmatic framework and model parameters etc. All the deep learning models have been implemented using the Python programming language by primarily relying on the *Transformers* library from Huggingface (Wolf et al., 2019). By using this library, it was possible to use pre-trained deep learning models, which limits the computational costs associated with training a model from scratch. Moreover, the library also provides tokenizers that are specifically suited towards the pre-trained models, which ensures that the input is transformed into the correct representation for a specific model.

The different models include multiple hyperparameters, which primarily have been tuned using a manual approach. Besides trying out multiple variations of hyperparameters, the main focus has more been on reusing parameters from related academic work, which have led to strong performances. As a result, all the models introduced in the subsequent sections has been fine-tuned using 5 epochs and a learning rate of $2e-5$. The input data is represented in batches with a size of either 128 or 256. Additionally, 1% of the total training steps has been used to warm-up the learning rate. The motivation for using a warm-up stage is to lower the primacy effect of the initial training instances (Hasty visionAI Wiki,

2021). For all models, the AdamW optimizer was used (Loshchilov and Hutter, 2017).

To constantly monitor the performance of the model training, the models are evaluated two times per epoch (10 evaluations in total) on a validation set, and the results are logged using the tool called Weights & Biases[2]. This tool makes it possible to monitor multiple metrics such as F1-score and loss, while the training is still running. Moreover, it also saves these evaluations so comparisons can be made based on multiple training rounds. When the training phase is completed, the best performing model is saved to Weights & Biases as well such that it can easily be applied on unseen data.

COMPUTATIONAL RESOURCES Training deep learning models with large amounts of text data constitutes a heavy computational task, which requires additional computing power to reduce training time. In this thesis, a single NVIDIA T4 Tensor Core GPU has been provided by Copenhagen Business School through UCloud, which is a platform for provisioning computational resources to researchers in Denmark. Furthermore, additional methods were also used to speed up the training process even further. To increase the speed of the model and limit memory usage, mixed precision was applied, which uses 16-bit floating-point types. Besides that, DeepSpeed[3] was also used to further improve the computational efficiency (Rajbhandari et al., 2019). As a result, the training time was reduced to a few hours instead of days.

4.5.3 Model 1: DistilBERT

The DistilBERT model introduced by Sanh et al. (2019) is a reduced version of the $\text{BERT}_{\text{base}}$ model from Devlin et al. (2018) that is 60% faster and retains 97% of the model's language understanding capabilities. More specifically, the DistilBERT model only constitute 60% of the size of BERT. As a result, the model is less expensive to pre-train and fine-tune, while still retaining most of BERTs capabilities. To both reduce the model size and maintain the same level of performance, Sanh et al. (2019) took advantage of knowledge distillation (Buciluǎ et al. 2006; Hinton et al. 2015) in the pre-training phase of the model. This is considered as a technique that can be used for compressing large models, where the goal

[2] https://wandb.ai/site
[3] https://www.deepspeed.ai/

is to achieve a solid model that can reproduce the capabilities of a larger model, which in this case is the BERT model. The general architecture of DistilBERT is similar to the one used by BERT. The customization has been done by discarding the pooler and token-type embeddings together with reducing the number of layers, which deemed to be the most beneficial changes to the modelling architecture. Furthermore, the initialization of the model is considered important in the training phase such that the sub-network is going to converge. As a result, only one layer of two from the BERT was used to perform the initialization of the model (Sanh et al., 2019). The data that was used to train BERT is also used for pre-training the DistilBERT model.

The motivation for applying this specific model is based on the computational benefits and the fact that the model is able to retain most of the original BERT model's capabilities. As a result, the model has primarily been used to perform initial experiments, which allowed for faster adjustments to optimize the performance of the model.

4.5.4 Analysis of Preliminary Results of Dataset Variations

In section 4.4.4, it was suggested to use an oversampled version of the combined training dataset. More specifically, it was only suggested to oversample the hate speech class because it was represented the least in the combined training dataset. In this section, the oversampled training set is tested against the non-oversampled training set to figure out the best composition of the training dataset for further experiments.

To conduct this experiment, the DistilBERT model is used based on its computational efficiencies. The results of DistilBERT fine-tuned with the normal and oversampled datasets are recorded in Table 22. Both datasets achieve effective results with a macro F1-score of around 80% on both the evaluation and the test sets. Oversampling turns out to be only marginally better compared to the normal dataset with a difference of between 0.1 and 0.2 percentage points. When analyzing class performance, it can be observed from the table that the oversampled training data performs better for both the normal and offensive class. However, the unsampled dataset performs slightly better on the hate speech class. As a result, all the other BERT models that are applied in the thesis is fine-tuned using the oversampled dataset.

Dataset	Label	n	D	D-OS
Train Set	Total	456,003	87.6	**87.9**
Eval Set	Total	47,261	80.0	**80.2**
Test Set	Normal	26,341	86.1	**86.2**
	Offensive	13,519	81.2	**81.4**
	Hate speech	7,042	**72.3**	72.2
	Total	47,262	79.9	**80.0**

D: DistilBERT, D-OS: DistilBERT (oversampled data)

Table 22: DistilBERT F1-score (%) for training, eval, and test datasets using both a non-oversampled and oversampled training dataset. Highest scores per dataset and label are marked bold.

To identify possible optimizations for deep learning models, one possibility is analyzing the error rates of a model. According to Ng et al. (2021), differences between train and evaluation datasets in a model's performance indicate a high variance in the training data. In this thesis, the oversampled dataset performs better on the train set, which is to be expected as samples appear multiple times in the dataset. When comparing train and eval as well as train and test, a drop in performance around 7.5% to 8% can be recorded. Due to the high variety and quality of some of the used train datasets, the variance drop is deemed acceptable, especially in the context of hate speech research.

Table 23 shows the accuracy on the test set for each dataset used in training. From the table, it is evident that are big discrepancies between the different datasets, ranging from over 90.8% for the `wikipedia` dataset to 52.7% for the `ousid` dataset. Generally, it can be observed from the results that the performance is higher for datasets that are represented the most in the training dataset. The reason for this can be attributed to the fact the distribution of the training dataset is reflected in the test data as a stratified split was performed, when splitting the dataset into train, validation, and test sets. Besides that, oversampling improved the performance on the datasets that were oversampled by up to 4.5%. For the datasets that were not oversampled some results were improved and others were worsened.

Label	n	D	D-OS
cad	113	64.6	**69.0**
civil	19,423	83.7	**83.7**
davidson	2,489	**88.1**	87.9
dynhs	4,085	72.0	**72.0**
ghc	2,803	80.9	**81.7**
hasoc	718	67.3	**69.5**
hatemoji	609	67.7	**68.3**
hateval	1,088	**71.1**	70.0
hatexplain	2,059	61.0	**62.9**
ousid	535	52.7	**53.8**
slur	3,890	**87.6**	87.5
wikipedia	9,450	90.7	**90.8**
Total	47,262	82.4	**82.5**

D: DistilBERT, D-OS: DistilBERT (oversampled data)

Table 23: DistilBERT model accuracy (%) per source dataset. Highest scores per dataset and label are marked bold.

4.5.5 Model 2: BERT

The BERT model introduced in section 2.4.3 primarily serves as a baseline such that it is possible to compare the different variations of BERT models that are being applied in this thesis. $\text{BERT}_{\text{base}}$ is considered both the first and most standard BERT model, and since it has been used as baseline in prior studies, it is considered a suitable baseline in this thesis as well, especially when using customized versions of BERT (Devlin et al., 2018). This baseline model is based on the encoding part of the transformer architecture from Vaswani et al. (2017), and it consists of 12 transformer blocks, a hidden size of 768, 12 self-attention heads, and 110M parameters. It is designed to handle a masked language modelling and next sentence prediction task, which makes it possible for the model to better understand language and context (Devlin et al., 2018). To be able to fine-tune the model for various classification tasks, an additional output layer is added on top of the model.

4.5.6 Model 3: RoBERTa

The Robustly optimized BERT approach (RoBERTa) was introduced by Liu et al. (2019) in a study that focused on replicating the pre-training phase of BERT (Devlin et al., 2018) in a way that multiple hyperparameters and data sizes were tested. Based on multiple experiments, Liu et al. (2019) stated that the original BERT was lacking training, and that further

performance gains could have been achieved if more training and other design choices were applied. As a result, Liu et al. (2019) proposed the RoBERTa model, which improves the pre-training phase of BERT.

On a high level, the model is trained for a longer time both using more data, larger batches, and longer sequences. The original BERT was trained using the BookCorpus (Zhu et al., 2015) and English Wikipedia corpus, which constitute 16GB of data. Contrarily, RoBERTa was trained on the same data together with three additional datasets, CC-News[4], OpenWebText (Gokaslan and Cohen, 2019), and Stories (Trinh and Le, 2018), which in total constitute 160GB of data. Multiple experiments were performed, and it was shown that when increasing the size and diversity of the training data, the performance increases. Regarding the batch size and number of training steps, the original BERT was trained with a batch size of 256 in 1M steps, whereas RoBERTa achieved better results using a batch size of 8K together with 500K training steps.

Additionally, the objective of the next sentence prediction task was removed together with applying a dynamic masking method. The original BERT relies on a static masking technique, where the masking is carried out a single time during the preprocessing step. This approach was compared with a dynamic masking technique that instead creates maskings when a sequence is inputted into the model. Multiple experiments were performed, and it was found that the dynamic masking technique achieved either slightly better or comparable performance to the static one in the experiments (Liu et al., 2019). Regarding the fine-tuning phase, RoBERTa uses the same procedures as the original BERT. Based on all these different adjustments, RoBERTa achieved new state-of-the-art results on multiple benchmark datasets such as GLUE (Wang et al., 2019), SQuaD (Rajpurkar et al. 2016, 2018), and RACE (Lai et al., 2017).

The motivation for using RoBERTa is to compare a model that is trained for a longer time and on much more data with both BERT and DistilBERT, which are considered as smaller models. This introduces an interesting comparison as it is possible to discuss whether a

[4] This data is collected by the authors themselves from the CommonCrawl News dataset (Nagel, 2016)

more robust model provides any improvements within the domain of hate speech detection. Furthermore, it also allows for a discussion regarding the advantages and disadvantages of using a larger model compared to smaller BERT models.

4.5.7 Model 4: HateBERT

The last model that is applied in this thesis is the so-called HateBERT language model (Caselli et al., 2021). This is a BERT model that has been re-trained using the Masked Language Model objective on RAL-E, a large-scale dataset of posts in English from banned communities on the social media site Reddit. As the model was introduced in section 3.3 as part of the literature review, it will not be explained here again.

The motivation for using this model is based on the fact that it has proved to be robust when modelling abusive language phenomena. According to Caselli et al. (2021), it was concluded that re-training the original BERT on abusive textual data makes it possible to increase the model's performance on this specific language variety even though it was not initially trained on abusive content.

4.6 FEATURE IMPORTANCE

Because of the inherit complexity of transformer-based deep learning models, they fall under the category of "black box" models. This means that it is between difficult and impossible for experts to explain how a model arrives at a specific decision[5]. This creates tension between accuracy and interpretability. In order to create trust in a model and to provide insight into improving a model , the ability to correctly interpret a prediction model's output is extremely important. To support this, methods are needed to explain the overall importance of a feature for a model (Lundberg et al., 2017).

SHapley Additive exPlanations (SHAP) is a unified framework for interpreting predictions from machine and deep learning models, based on cooperative game theory. It works by assigning each feature an importance value for a particular prediction. SHAP values are calculated by using fair allocation results from cooperative game theory to allocate credit

[5] Figure 21 shows a visualization of all attention layers of a BERT model on an example tweet from the UNHCR dataset.

for a model's output among its input features. The input properties of a model are matched to the players in a game, and the model function is matched to the game's rules. Using this, features can "join" or "not join" a model by using only a subset S of features. To evaluate a model f, the other features are integrated out using a conditional expected value formulation:

$$E[f(X)|do(X_S = x_s)] \tag{2}$$

In this formulation, the values of the features in S are known because they are set. This can be used to explain how a model would behave if one were to intervene and change its inputs. In the context of this thesis, features are tokens that get masked out from the sentence, which SHAP uses to observe the impact on the prediction score for each class.

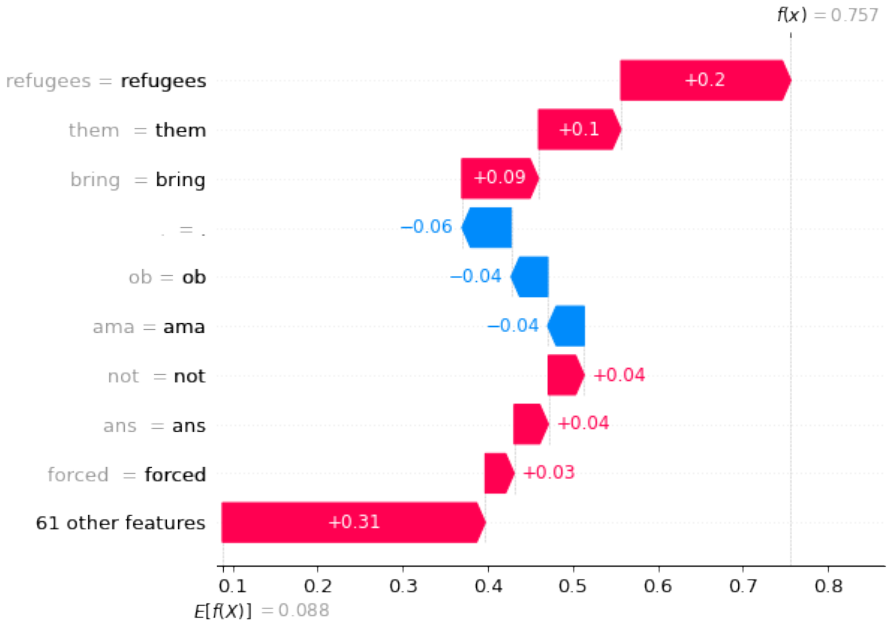

Figure 12: Waterfall plot showing feature importance for the hate class on one sample from the UNHCR dataset, which has been predicted by the RoBERTa Model.

Shapley values have the fundamental properties of always summing up to the difference between the game outcome when all players are present and the game outcome when no players are present. For the use on NLP models this means that SHAP values of all the input features, e.g. the full text without any masked tokens, will always sum up to the difference between the baseline (expected) model output and the current model output, e.g. confidence score for the prediction and class being explained (Lundberg et al., 2017).

Figure 12 shows this in a visualization using a waterfall plot that starts at the background prior $E[f(X)]$ for each class, in this case the hate speech class, and then adds tokens one at a time until the current model output $f(x)$ is reached (Lundberg, 2018). The force plot, used as visualization in this work is a collapsed and sorted version of the waterfall plot. In order to calculate the SHAP value ϕ_i for the same token x_i from multiple samples the arithmetic mean of all token x_i is used.

$$\sum_{c=1}^{C} \phi_i^c = 0 \tag{3}$$

Additionally, equation 3 shows that for any token x_i the sum of SHAP values for all classes C equals zero. This is also true when the mean of multiple samples is used. By using this method, which is provided in a Python package, it is possible to analyze samples of different datasets to extract the importance of tokens and to compare their impact on different classes.

4.7 EVALUATION

In the "Evaluation" phase of CRISP-DM, the proposed models from the "Modelling" phase are evaluated. Therefore, this section introduces the various evaluation strategies that are used in this thesis. Detecting online hate and abusive speech are difficult tasks for state-of-the-art models. As a result, it is not enough to rely on singular measures to evaluate the models. Instead, the goal is to allow for an identification of not only the best model, but also of weak points where the models are struggling to classify hate and offensive speech. Therefore, multiple evaluation strategies and metrics are proposed in this section.

First, an in-dataset evaluation is carried out using a test set from the combined dataset to test the models ability to detect hate and offensive speech from a variety of data sources using the metrics defined in the subsequent section. Furthermore, to test the models' ability to detect hate speech in the context of refugees, the hate speech dataset that was obtained from UNHCR is used as a test set using similar metrics. Besides that, to investigate the weak points of a model, several functional tests from HATECHECK (Röttger et al., 2020) were carried out. This is further explained in section 4.7.2.

4.7.1 Metrics

Common for most of the evaluation experiments presented above, an accuracy, precision, recall, and F1-score were used to perform the statistical evaluation of the models. The formulas in this section uses abbreviations such as TP, TN, FP, and FN. The explanation of these abbreviations is explained in Table 24.

Abbreviation	Name	Definition
TP	True Positives	An outcome where the model correctly predicts the positive class
TN	True Negatives	An outcome where the model correctly predicts the negative class
FP	False Positives	An outcome where the model incorrectly predicts the positive class
FN	False Negatives	An outcome where the model incorrectly predicts the negative class

Table 24: Explanations of true positives, true negatives, false positives, and false negatives. The definitions are obtained from Google (2020).

ACCURACY DeepAI (2022) defines this metric as "*the number of correctly predicted data points out of all the data points*", and it is considered as one of the most commonly used metrics for evaluating classification models. The formula for calculating this metric is presented in Equation 4.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

The accuracy metric contains limitations when working with imbalanced datasets. The reason for this is that if a high accuracy is achieved, it does not necessarily mean that the model is performing well. The high accuracy can be a result of the imbalanced dataset because if the model is able to perfectly classify the majority class, then it will lead to a high accuracy. If the model is only able to classify a single instance of the minority class correctly, then this is overshadowed by the performance on the majority class.

F1-SCORE The F1-score can be defined as "*the harmonic mean of precision and recall*" (Géron, 2017), and constitutes a better metric when working with an imbalanced-dataset, which is the case in this thesis. The benefit of using this metric is that it adds a higher weight to low values (Géron, 2017). The formula is presented in Equation 7, but in order to understand the formula, the two metrics, precision and recall, will be defined first as they constitute most of the formula.

The **precision** metric aims at answering "*what proportion of positive identifications was actually correct?*" (Google, 2020). The formula is presented in Equation 5.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

On the other hand, the **recall** metric aims at answering "*What proportion of actual positives was identified correctly?*" (Google, 2020). The formula is available in Equation 6

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

When working with both the precision and recall, it is important to be aware of the **precision/recall tradeoff**. This means that when the precision is rising, the recall is reduced, and vice versa. Now, that both precision and recall have been defined, it is possible to construct the formula for the F1-score, which is presented in Equation 7.

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

The F1-score is not a single metric as there are multiple variations of it, which depends on whether the classification task at hand is considered binary, multi-class, or multi-label. To evaluate the multi-class experiments, this thesis mainly relies on the **macro F1-score**. Prior related academic work primarily used the macro F1-score, and therefore, it is also used in this thesis because it makes it possible to make a comparison.

When the averaging strategy is set to macro, it means that the F1-score is calculated for each of the classes. Afterwards, the unweighted mean is calculated based on the results for each class (scikit-learn, 2022). Using this specific metric, makes it possible to treat each class equally as the same importance is assigned to each of the classes (PELTARION, 2022). The formula for this variation of the F1-score is presented in Equation 8, where N is the number of classes.

$$\text{Macro } F_1 = \frac{1}{N} \sum_{i=0}^{N} F_1 - \text{Score}_i \tag{8}$$

4.7.2 HATECHECK

Only evaluating on metrics as F1-score or accuracy risks overestimating the generalisability of a hate speech model. To enable more targeted diagnostic insights, HATECHECK is applied to better highlight areas where the model exhibits weaknesses. HATECHECK is a dataset provided by Röttger et al. (2020) that consists of multiple functional tests that can be used to validate hate speech detection classifiers. The concept of functional tests originates from software engineering, and Ribeiro et al. (2020) showed its usefulness in various natural language processing tasks. Röttger et al. (2020), subsequently presented this framework specifically for the area of hate speech detection. Yet, HATECHECK can merely demonstrate the lack of a specific flaw, rather than necessarily describing a generalizable model strength. The framework, therefore, provides focused diagnostic insights as a supplement to, rather than as a substitute for evaluating real-world hate speech on held-out test sets.

The framework contains 29 functional tests in total, distributed over 3,728 labelled instances, with 18 testing for various forms of hateful content and distinct expressions of hate. The other 11 tests focus on non-hateful contrasts. To construct the functional tests, the authors interviewed 21 NGO workers and used previous findings from the literature. Besides that, the authors provide overall 866 different templates, which contain the `[IDENTITY]` placeholder. This placeholder then gets replaced by the protected groups used in the paper; woman, trans people, gay people, black people, disabled people, Muslims and immigrants (Röttger et al., 2020). Not all of the 7 protected groups are used in all test cases, due to some of them using specific slurs or specifically misspelled words. The immigrant group is of special interest to this thesis because the group is similar to refugees. Furthermore, to increase the usability of these tests in the context of refugees, the HATECHECK datasets is expanded in this thesis with 421 additional test cases for the identity "refugees". This is integrated for test cases 1-20. The different functional tests together with associated examples from HATECHECK can be found in Table 31.

To summarize this chapter, multiple datasets were collected and prepared. More specifically, 12 different datasets from various sources that have been used to detect hate and offensive speech in a variety of contexts were combined into one large dataset. This dataset

primarily acts as a training set for the models, but a subset of the data is used to evaluate the in-dataset performance. Besides that, a hate speech dataset targeted at refugees was provided by UNHCR such that the models' ability to detect hate surrounding refugees can be evaluated. Furthermore, five unlabeled datasets were obtained from Wolters and Olšavský (2021) such that it is possible to analyze the change of the amount of hate, offensive, and normal speech around specific international refugee crises over time. To perform the experiments, it was decided to use the following four models; DistilBERT, BERT, RoBERTa, and HateBERT. Moreover, for evaluating the models' performance, multiple evaluation metrics was introduced, where the accuracy, F1-score, and HateCheck (Röttger et al., 2020) is primarily going to be used.

5 RESULTS

The results of the experiments introduced in chapter 4 are presented in this chapter using the evaluation metrics from section 4.7.1. First, the in-dataset experiments are evaluated, where the focus is to analyze how robust the models are on data from the same domains as they are trained on. Subsequently, the models' ability to detect hate speech in the context of refugees will be tested using refugee-related data provided by UNHCR, which has not been seen by the models before. Furthermore, an error analysis is also conducted using the HATECHECK dataset such that it is possible to identify specific scenarios where the models are struggling and can be improved. After evaluating the models' performance on multiple datasets, the best model is applied to unlabeled international refugee crisis data in order to give an indication of how hate, offensive, and normal speech can be monitored over time. Furthermore, a comparison between the crises is made.

5.1 IN-DATASET EVALUATION

To identify the best performing model, the predictions of each model on the test set are evaluated using the macro F1-score, which is presented in Table 25. The RoBERTa model has the highest total F1-score with 81.0% and a lead of 0.3 percentage points over the baseline BERT model.

When comparing the overall per-class performance, it can be observed that there is a large difference in the performance on the normal and hate speech class for all the models. This

Label	n	DistilBERT	BERT	HateBERT	RoBERTa
Normal	26,341	86.2	**86.9**	86.7	86.8
Offensive	13,519	81.4	81.7	81.6	**81.9**
Hate speech	7,042	72.2	73.6	73.7	**74.2**
Total	47,262	80.0	80.7	80.6	**81.0**

Table 25: Model F1-scores (%) per class on the test dataset. Highest scores per class are bold.

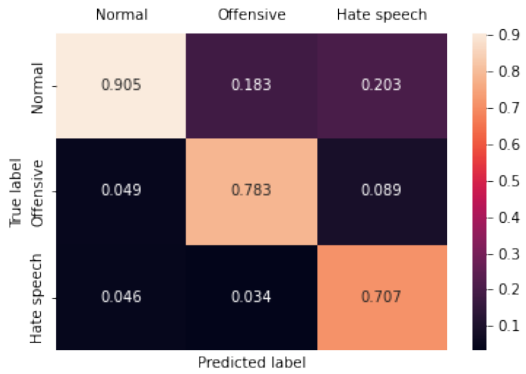

Figure 13: Confusion matrix of RoBERTa model. The values are normalized for each of the predicted labels.

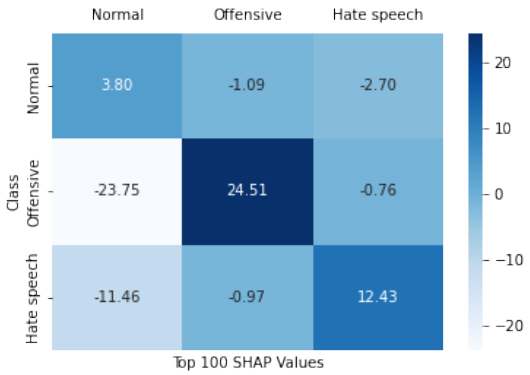

Figure 14: Feature importance of top 100 tokens (columns) for each class using RoBERTa. Displayed is the sum of SHAP values for each class (rows).

can especially be observed for the DistilBERT model, where the difference is 14 percentage points. Generally, the DistilBERT model is outperformed by all the other models on all classes. Besides that, the RoBERTa model reaches the highest F1-score on the hate speech class with 74.2%, followed by the HateBERT model with 73.1%. From the table, it is also evident that RoBERTa outperforms the other models on the offensive class only by a small margin. What is interesting, based on the results, is that the baseline BERT model outperforms RoBERTa on the normal class even though RoBERTa has been pre-trained using a larger amount of general text data. However, it is only with a lead of 0.1 percentage points. When focusing more on HateBERT, it can be observed that it performs a little better than the baseline BERT model on the hate speech class with a score that is 0.1% higher.

Figure 13 shows the confusion matrix for the best performing model, RoBERTa, normalized for the predicted label. This means that all predictions are being displayed as a percentage of true labels. Therefore, the True Positives TP for each class is the accuracy. The model performs best on the normal class with an accuracy of 90.5%. Besides that, the misclassifications are distributed equally between the two other classes. Furthermore, it can also be observed that when the model predicts text as being hateful, the true class is offensive in 8.9% of the cases. On the other hand, when the model predicts text as being offensive, the true label is hate speech in only 3.4% of the cases. These misclassification rates are deemed acceptable, especially when considering that even humans struggle to distinguish between

these two classes in some scenarios. Most of the time both the hate speech and offensive class are misclassified as normal.

FEATURE IMPORTANCE To calculate feature importance for RoBERTa, the SHAP values introduced in section 4.6 are used. Figure 14 shows the cumulative contribution of the top 100 tokens for each class (column) on itself and on the other classes (row). This indicates the prediction influence of the most important features. The scores are derived from the mean SHAP values of the top 100 most important features, as displayed in Table 26, which are summed up for each class. Remarkably, the top 100 most important tokens of the offensive class have with 24.51 twice as much impact on itself as the hate speech class with 12.43. Besides that, it can also be observed from the figure that the top offensive and hate speech features have little impact on each other. This is somewhat similar to the confusion matrix. The low scores for the normal class on itself suggest that the absence of hateful and offensive tokens in combination with many tokens of low feature importance contribute towards the model classifying a sample as normal.

Token	n	N	O	HS
igger	13	−0.27	−0.13	0.40
gays	4	−0.30	−0.08	0.38
agg	69	−0.22	−0.12	0.34
sexual	4	−0.26	−0.04	0.30
anny·	46	−0.24	−0.05	0.30
trans·	15	−0.22	−0.07	0.29
igger·	41	−0.21	−0.08	0.29
lems·	4	−0.23	−0.04	0.27
ays·	4	−0.25	−0.01	0.26
queer·	4	−0.14	−0.12	0.26
ike·	8	−0.18	−0.07	0.25
aliens·	5	−0.20	−0.05	0.25
blacks·	8	−0.17	−0.04	0.21
immigrant·	7	−0.16	−0.04	0.20

(a) Hate speech

Token	n	N	O	HS
stupidity·	8	−0.51	0.55	−0.04
pussy·	12	−0.51	0.53	−0.02
ridiculous	6	−0.51	0.53	−0.02
bitch·	28	−0.49	0.52	−0.02
pussy	6	−0.44	0.47	−0.03
ches·	19	−0.44	0.47	−0.02
stupidity	4	−0.46	0.47	−0.01
fool·	7	−0.44	0.46	−0.02
stupid	23	−0.44	0.46	−0.02
oes	4	−0.38	0.43	−0.05
ches	4	−0.53	0.43	0.10
loser·	8	−0.40	0.42	−0.02
hypocr	8	−0.37	0.39	−0.02
fools·	4	−0.35	0.39	−0.04

(b) Offensive speech

Mean SHAP values for class N: Normal, O: Offensive, HS: Hate speech

Table 26: Feature importance of top tokens with more than 3 occurrences sorted by mean SHAP value for class a) hate speech and b) offensive speech. SHAP values were calculated on a subset of 2000 samples. Whitespace after a token is marked with a (·).

Table 26 shows the 14 tokens with the highest feature importance for the test set on the hate and offensive class. Only tokens that are present three or more times are shown to exclude outliers and increase generalizability. Additionally, tokens that end with a whitespace have been marked with a middle dot (·). This indicates that the token is followed by another word and not by another token of the same word or punctuation. Since RoBERTa uses byte-level Byte-Pair-Encoding for tokenizing words, the following tokens; "*agg*", "*ays*", and "*oes*" can be mapped to their original words by referencing Table 36 in the Appendix. From the tables, it can be observed that tokens such as "*igger*", "*gays*", and "*agg*" are most influential for the hate speech class. On the other hand, tokens such as "*stupidity·*", "*pussy·*", and "*ridiculous*" are most important for the offensive class. For both classes, it is clear that these top 3 most important tokens have a negative contribution towards the other classes. This finding applies to most of the tokens on the lists, with the "*ches*" token (offensive) being a notable expection as it also increases the likelihood of hate speech by 0.10.

5.2 REFUGEE-RELATED DATA EVALUATION

Next, the models performance was evaluated using test data provided by UNHCR, which exclusively contains hateful tweets against refugees. The result for each of the models is presented in Table 27. Since this test dataset only consists of 98 examples of one single class, the accuracy metric is used to evaluate the performance of the models.

Label	n	DistilBERT	BERT	HateBERT	RoBERTa
Hate speech	98	64.3	71.4	71.4	**73.5**

Table 27: Model accuracy (%) on refugee test data. Highest-score is marked bold.

From the table, it can be observed that most of the models obtain a similar accuracy score, except the DistilBERT model, which is lacking behind with a score of 64.3%. When focusing on both the BERT and HateBERT models, it is interesting that a model like HateBERT, which is a regular BERT model that is re-trained on hateful content, is not achieving a better performance than BERT. This suggests that the re-training with hateful text is not having an effect on this test dataset. A reason for this can possibly be attributed to the data that has been used to re-train BERT. If the hateful terms and phrases in that dataset are completely different from what is present in the UNHCR refugee data, the re-training will

not constitute an advantage. The size of the test dataset can also have an influence on the performance of the models as there is only 98 tweets to evaluate the models on.

The best performing model is RoBERTa, which achieves an accuracy score of 73.5% and outperforms the baseline model (BERT) with 2.1 percentage points. This indicates that it is more beneficial to use a model that is both pre-trained for a longer time and pre-trained using much more data than the other models for this classification task (Liu et al., 2019).

ERROR ANALYSIS This section provides an analysis of a subset of misclassified tweets from the RoBERTa model as it was identified above that even though it is the best performing model, which achieved an accuracy of 73.5%, it is still not perfect. Table 28 shows a subset of misclassifications, where the model has predicted the tweets as not containing hate speech. Since the model was initially trained to classify text as either being hate speech, offensive, or normal, the misclassifications constitute text instances that are either classified as offensive or normal. More specifically, 3 text instances were classified as being offensive, whereas 23 were classified as being normal. Furthermore, the table also contains the confidence scores of the model to indicate how confident the RoBERTa model was when making this prediction.

Example ID	Misclassified Example	Prediction	Confidence Score		
			Normal	Offensive	Hate speech
1	"@USER @USER turkey welcomes more than 5 million syrian refugees, remind me when france opens its borders at only 1 million dirty puffon"	Offensive	0.15	**0.45**	0.40
2	"@USER @USER granting asylum to refugees fleeing from oppression who are members of salafist terrorist groups has consequences, particularly if you allow them to raise their sons with their rancid ideology. clear enough, or do you require fewer syllables?"	Normal	**0.70**	0.00	0.30
3	"there are no refugees, venezuelan leaves itself. ; why they use the word "refugees" when they mention venezuelan migrants. if there is not a war in venezuela. and according to mature is the best country to live in latin america, the zurderio is pisa the tail migrants venucas"	Normal	**0.91**	0.00	0.08
4	"syrians that translate the country to asylum-seeker dumps ... are turning to the largest criminal crime organization."	Normal	**0.46**	0.25	0.29

Table 28: Sample of misclassified tweets using RoBERTa. Highest scores are marked bold.

From the table, it can be observed that the model was not very confident with its prediction (offensive) for example 1. By looking at the confidence scores for each class, it is clear that the model had a difficult time deciding whether this particular tweet is offensive or hate

speech. When comparing the tweet with the hate speech definition provided by the UN (see section 2.1), it can be observed that this tweet does not fall into the category of hate speech. Therefore, this misclassification might be a result of an annotation error. Instead, the tweet points more into the direction of being offensive because of the phrase "*dirty puffon*", which is considered offensive by the authors of this thesis. However, the rest of the tweet does not contain any signs of either offensive or hate speech.

When focusing on example 2, it is clear that the model is more confident than for the first example, but it is still not completely sure. The reason for this misclassification can be attributed to the fact that the tweet is lacking context as it is not clear whether it is meant to be informative or hate speech. Therefore, it is also difficult for a human to assess whether it is hate speech or not without having an understanding of the discussion thread the tweet appears in. Additionally, it seems that some of the words in the tweet have a high weight toward the hate speech class as the model indicates that there is 30% chance that it is hate.

It can be observed that example 3 does not contain any indications of hate speech. As a result, it is considered an error associated with the annotation of the data, which was handled by UNHCR. Moreover, the confidence score also shows that the model is able to clearly identify a tweet that does not contain any signs of hate speech. Regarding example 4, the model is again a bit challenged due to its low confidence score for all classes. By analyzing the tweet, it is clear that it contains hate speech as the sender accuses the Syrian refugees for "*turning to the largest criminal crime organization*". A reason why the model has misclassified this tweet could be explained by the fact that the tweet does not contain any words or phrases that include either slurs or profanity, which are often found in hate speech. One thing that is important to mention regarding all these misclassifications is that some of them have been translated into English by UNHCR. As a result, there is a chance that some of the tweets have lost context, which confuses the model when it has to predict.

FEATURE IMPORTANCE This section gives some insights into the most important features for the RoBERTa model, when it has to classify whether a tweet is considered hate speech or not. First, the most important features for the whole test dataset will be analyzed

to see which tokens have the highest influence on the predictions. Afterwards, a specific tweet will be analyzed to see which tokens steer the direction of the prediction. Since this test dataset only consists of hateful examples, it is only the most important features for the hate speech class that are considered.

In Figure 15, the 12 most important features for the hate speech class are visualized. Here, the mean SHAP value for each feature is used to represent the overall importance. Since RoBERTa performs byte-level Byte-Pair-Encoding for tokenizing the words that are not present in the vocabulary, some of the tokens in the visualization only constitute a part of the original word. For example, the most important feature in this specific test data is "*abs*", which is derived from the word "*arabs.*". In Table 36 in the Appendix, these tokens are mapped to their original words. Furthermore, when interpreting the values in Figure 15, it is important to remember that the importance of a specific feature depends on the contexts that it appears in, which is the benefit of transformer-based models. Therefore, it is not sure that a specific feature has the same importance in other contexts.

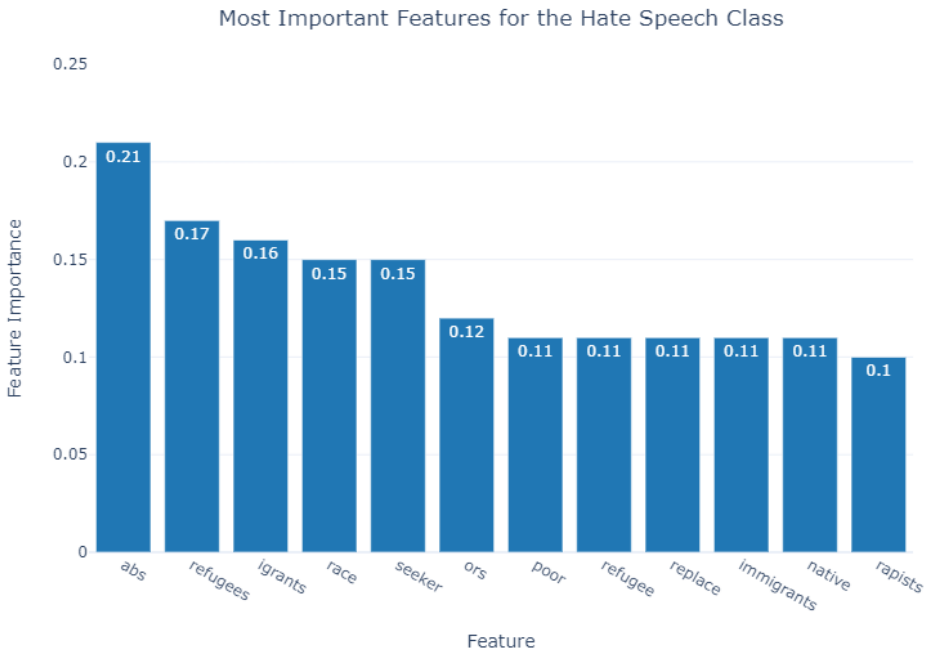

Figure 15: Most important features for predicting hate speech on the refugee test dataset using the RoBERTa model.

From the figure, it can also be observed that tokens such as "*refugees*", "*igrants*", "*refugee*", and "*immigrants*" appear to be among the most influential features. This is in line with UNHCR's findings because they have identified that if these tokens appear in a tweet, it is most often associated with hateful content. Besides that, the tokens "*poor*" and "*rapists*", which are often used in relation with accusations, also have a minor influence when classifying tweets as hate speech using this specific test dataset.

Supplementary to the 12 most important features, a tweet-level analysis is provided in Figure 16 to understand how the RoBERTa model made its classification. Here, it is possible to observe which tokens were most influential on a tweet-level. The tweet that is presented in the figure is correctly classified by the RoBERTa model with a confidence score of 99%. The red color indicates that a token is contributing to the hate speech class in this specific context, whereas the blue color represents tokens that support the opposite class. The saturation of the color shows the importance of each specific token, which means the higher the saturation, the more influential a token is. In the figure, it is difficult to observe the importance score for each feature because it is only possible to get this value by hovering the visualization in a Jupyter Notebook or HTML format. As a result, it is only important to focus on the part where the tokens are highlighted with color saturation.

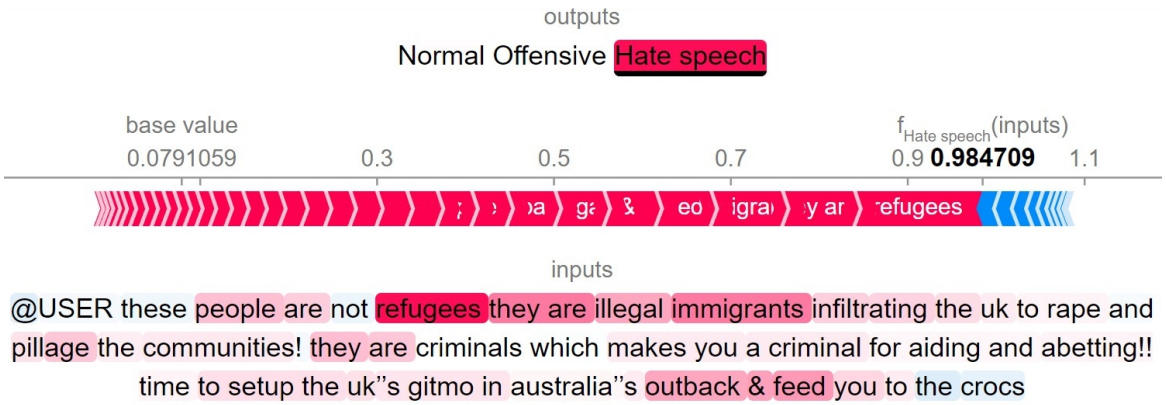

Figure 16: The influence of each feature in a tweet is represented using a force diagram showing SHAP values to analyze how the RoBERTa model correctly classified the tweet as being hate speech. The red tokens contribute to hate speech class, whereas the blue tokens has a negative influence. This means that these tokens contribute to the opposite classes.

From Figure 16, it is evident that there are multiple features that support the hate speech class in this specific context. The token "*refugees*" is the most influential token with a SHAP

value of 0.15, when considering each feature separately. This finding is in line with what was identified in Figure 15. Besides that, the token "*immigrants*" and the phrase "*they are*" are also considered as highly influential features in this specific context. Furthermore, it is also interesting to see that since the token "*people*" is considered important in this context, the phrase "*they are*", which refers to "*people*", also has an influence. This indicates that the model is able to understand that the phrase "*they are*" refers to "*people*".

By performing this type of analysis, it is possible to identify which features influence the classification and by how much. Also, when calculating the importance score for each feature using the whole dataset, then it is only represented as a mean value. Instead, when performing the tweet-level analysis, the importance for a feature in a specific tweet can be deduced.

5.3 HATECHECK FUNCTIONAL TESTS

To further evaluate the models, this section focuses on identifying weaknesses associated with the models using the HATECHECK dataset from Röttger et al. (2020), which is introduced in section 4.7.2. This analysis is run on the baseline model (BERT), HateBERT, and RoBERTa. The DistilBERT model is not included here because it was completely outperformed by the other models on the refugee-related data in section 5.2. Before diving into a more fine-grained analysis, the overall performance on the dataset is evaluated to see whether the models perform better on text that contains hate speech versus text without hate speech. More specifically, the accuracy metric is used to evaluate the models in this section as it was also used by the authors of HATECHECK (Röttger et al., 2020). This makes the results more comparable with other academic work. Furthermore, the highest accuracy score across the different models are highlighted with **bold** in the tables.

Since the models in this thesis are trained to classify text as being either hate speech, offensive or normal, both the offensive and normal predictions in this section will be treated as the non-hateful class because HATECHECK constitutes a binary classification task. Importantly, HATECHECK can only be used to discover the areas where the models are struggling, not where they perform well. This means that if a model performs well in a certain test, it

is only possible to conclude the absence of a weakness. Therefore, one should be careful with characterising generalizable strengths from HATECHECK (Vidgen et al., 2020).

Label	n	BERT	HateBERT	RoBERTa
Hateful	1,261	89.3	93.7	**94.0**
Non-hateful	2,888	67.6	72.6	**81.9**
Total	4,149	82.7	87.2	**90.3**

Table 29: Model accuracy (%) on HATECHECK data by test case label. Highest scores are bold.

OVERALL PERFORMANCE Table 29 presents the results of this binary classification task, and it can be observed that RoBERTa outperforms HateBERT and the baseline model, BERT, both on hateful and non-hateful text samples. More specifically, RoBERTa obtained an accuracy score of 90.3%, which is 3.1 percentage points higher than HateBERT and 7.6 percentage points higher than the baseline. When focusing on the hateful class, RoBERTa outperforms HateBERT and the baseline with a margin of 0.3 and 4.7 percentage points respectively. What is more interesting is the fact that the performance on the non-hateful class drops significantly for each of the models. RoBERTa achieved an accuracy score that is 12.1 percentage points lower whereas HateBERT and the baseline model achieved an accuracy score that is 21.1 and 21.7 percentage points lower respectively than what it achieved on the hateful data. Even though, all the three models experience a drop in performance on the non-hateful text, it is interesting that the reduction in the accuracy score is significantly less for RoBERTa than the other models.

Based on these results, it can be argued that all the models present some weaknesses in relation to the non-hateful class, but to get a more clear picture of this, a more fine-grained analysis is needed to figure out specifically where the models are struggling. One reason for the drop in performance can be associated with the fact that the non-hateful text examples are specifically constructed to be hard cases to challenge the models. In some of the hateful sentences, only one word is either removed or substituted such that it becomes non-hateful by still preserving almost the same grammatical structure. This can be seen in the following sentence "*Immigrants do not deserve to be tortured like that.*" where the phrase "*do not*" is added to make it non-hateful.

The results also suggest that a more robust BERT model, which is pre-trained on more general text data performs better on the HATECHECK dataset than a more specific BERT model that is re-trained using abusive content. Both models obtain a similar result for the hateful class, but the benefit of a more general pre-training strategy becomes visible when focusing on the non-hateful class. Based on these results, it is now possible to compare the results with other academic work, which has used HATECHECK to evaluate their models as well. It is important to mention that the authors of this thesis added additional test case for the *Refugees* target group, which means that the models have been tested on a modified HATECHECK dataset. As a result, it is not possible to make a 100% valid comparison, but since the models were run on the original HATECHECK dataset before modifying it and similar results were obtained on the modified version, the authors of this thesis allow themselves to make this comparison anyway.

The RoBERTa model outperforms the best model that was applied by the creators of the HATECHECK dataset (Röttger et al., 2020) with a margin of 13.7 percentage points. More specifically, Röttger et al. (2020) obtained their best result, 76.6%, using the commercial model, Perspective, which is developed by Google Jigsaw. The accuracy scores for both the hateful and non-hateful classes are 89.5 and 48.2 respectively. Furthermore, it is also interesting to compare with the results from Vidgen et al. (2020) because the authors used a similar model as in this thesis. More specifically, a $\text{RoBERTa}_{\text{base}}$ model from Huggingface was used, which obtained an accuracy score of 95%. Moreover, the model achieved a 95% and 93% accuracy score for the hateful and non-hateful classes respectively. As a result, Vidgen et al. (2020) outperform the best model in this thesis. One reason for this can possibly be explained by the performance on the non-hateful class. Here, Vidgen et al. (2020) obtains an accuracy score that is 11.1 percentage points higher, which suggests that the model applied in this thesis contains some weak spots regarding this type of data. This further indicates that the training data used by Vidgen et al. (2020) is well suited to both handle the hateful and non-hateful examples in the HATECHECK dataset.

TARGET GROUPS To perform a more fine-grained analysis of the performance of the models, the predictions can further be split into target groups. This makes it possible to identify whether the models show any weaknesses towards specific target groups. From Table 30, it is evident that even though all three models obtain relatively high accuracy scores, RoBERTa and HateBERT still outperform the baseline model (BERT) by a significant margin. This means that both the RoBERTa and HateBERT models are performing better on all the target groups. Moreover, RoBERTa is also outperforming HateBERT as the model achieved a higher accuracy score for each of the target groups. As a result, RoBERTa will be the center of this discussion.

Target Group	n	BERT	HateBERT	RoBERTa
Black people	421	80.9	84.0	**89.8**
Disabled people	421	80.4	85.5	**90.1**
Gay people	421	81.9	88.4	**88.9**
Immigrants	421	81.9	88.1	**92.2**
Refugees	421	83.1	89.5	**91.0**
Muslims	421	83.9	87.6	**90.3**
Trans people	421	85.5	88.3	**92.7**
Women	421	74.7	80.2	**83.1**

Table 30: Model accuracy (%) on test cases generated from `[IDENTITY]` templates by targeted protected group. Highest scores are marked bold.

When looking more specifically at the results, it can be observed that the model shows weaknesses toward the "Women" target group as the accuracy score of 83.1% is significantly lower than the rest of the scores. A reason for this can possibly be attributed to the datasets that were used for fine-tuning the model because if the textual content surrounding women is underrepresented, it would potentially be reflected in these results. Besides focusing on the worst performance, it can also be observed that RoBERTa performs best on the "Trans people", "Immigrant", and "Refugees" target groups, where the model achieved 92.7%, 92.2%, and 91.0% respectively in accuracy. Especially the results for the "Immigrant", and "Refugees" target groups are highly appreciated as the main focus of this thesis is hate speech detection against refugees. Also, these terms are often both used in the same context and interchangeably, which makes the results highly relevant.

FUNCTIONAL TESTS To perform an even more fine-grained analysis of the models' weaknesses, the HATECHECK dataset consists of 29 functional tests that makes it possible to more specifically identify the areas where the models are struggling. As a result, it can be used to figure out where to optimize the models. Table 31 presents the results of all three models for each of the 29 functional tests. Besides that, the table also includes example sentences to support the readers understanding of the different functional tests. From the table, it can be observed that the models overall achieve high accuracy scores for most of the functional test, which is considered positive.

What is more interesting, and also the purpose of using HATECHECK is to analyze, which types of tests the models are struggling with. From the table, it is clear that all three models are failing the most in **F15**, which includes sentences that contain negations. More specifically, it is "*non-hate expressed using **negated** hateful sentences*". Both the BERT and HateBERT are performing worse than a random baseline as the accuracy scores are lower than 50%. On the other hand, RoBERTa achieves an accuracy of 57.9%, which is close to exhibiting a random behaviour. The example test cases for this functional test are considered hard because it is only one or two words that separate the hateful and non-hateful cases.

When focusing on **F20** and **F21**, it can also be observed that the models are having a difficult time handling counter speech. Both the baseline model and HateBERT achieve accuracy scores that are around 50%, which indicates that the two models are failing to handle these types of text instances. On the other hand, RoBERTa is much better at handling counter speech since it achieves 79.5% and 73.1% in accuracy for **F20** and **F21** respectively. Furthermore, the models also seem to have some difficulties regarding non-hateful text against protected group identifiers (**F18** and **F19**) as most of the accuracy scores are in the range of $70\% - 80\%$. However, it is interesting that RoBERTa completely outperforms the other models on **F19**, which consist of "*positive statements using protected group identifiers*".

When focusing on the functional tests that includes slur usage (**F7**, **F8**, and **F9**), it can be observed that HateBERT overall outperforms the other two models. The reason for this can possibly be attributed to the fact that HateBERT has been re-trained using an abusive

text dataset before being fine-tuned, where there is a high chance that multiple slurs was present. One interesting thing is that both HateBERT and RoBERTa are performing much worse than the baseline on **F8**, which consists of "*non-hateful homonyms*". This is the only test where this behaviour is observed. Moreover, the results also indicate that the models can be improved in this area as the performance is not on the same level as most of the other high-performing categories. Lastly, the benefits of using a model that is pre-trained on larger amounts of general data and for a longer time is expressed when focusing on text instances that include spelling variations. Here, RoBERTa overall outperforms the other two models, which is expected when the model is pre-training using large amounts of general text data.

To summarize the findings of this analysis, it can be stated that all three models contain weaknesses in different areas, therefore, there are room for improvements. More specifically, it was identified that all models exhibits a weakness towards the non-hateful class in this dataset, which is possibly caused by the presence of really hard cases. The reason why these cases are considered hard is because the linguistic parts of the sentences are similar for both hateful and non-hateful text. Besides that, when analysing the models' behaviour towards target groups, it was shown that all the models had a more difficult time dealing with texts that were targeting women. The reason for this was that the accuracy score was significantly lower than for the other target groups. Furthermore, when performing the 29 functional test, it was identified that all models exhibited larger weaknesses regarding texts that included negations. A similar behaviour was discovered for the baseline model and HateBERT on counter speech. Overall, when looking at all the results for the functional test, it can be stated that both RoBERTa and BERT obtain higher accuracy scores for tests that included non-hateful instances, which is expected since both models are pre-trained on general text data. On the other hand, HateBERT more often achieved higher accuracy scores on hateful content than the two other models. This shows the benefit of re-training a BERT model with abusive text.

	Functionality	Example Test Case	Gold Label	n	Accuracy (%) B	H	R
Derogation	**F1**: Expression of strong negative emotions (explicit)	"I hate [IDENTITY]." "You [IDENTITY] make me sick."	hateful	160	98.1	**99.4**	**99.4**
	F2: Description using very negative attributes (explicit)	"[IDENTITY] are disgusting." "All [IDENTITY] like them are vile."	hateful	160	98.1	**99.4**	98.1
	F3: Dehumanisation (explicit)	"[IDENTITY] are scum." "[IDENTITY] are a pest."	hateful	160	98.1	**99.4**	98.8
	F4: Implicit derogation	"[IDENTITY] belong in a zoo." "I'd rather die than date [IDENTITY].	hateful	160	95.6	**99.4**	97.5
Threatening language	**F5**: Direct threat	"I am going to shoot the next [IDENTITY] I see."	hateful	152	84.2	**93.4**	92.8
	F6: Threat as normative statement	"We should get rid of [IDENTITY] once and for all."	hateful	160	91.2	**95.0**	**95.0**
Slur usage	**F7**: Hate expressed using slur	"You are just a [SLUR] to me."	hateful	144	77.1	**84.7**	74.3
	F8: Non-hateful homonyms	"I had this queer feeling we were being watched."	non-hate	30	**96.7**	86.7	80.0
	F9: Reclaimed slurs	"I'm the freshest n*gga in the room."	non-hate	81	77.8	84.0	**86.4**
Profanity usage	**F10**: Hate expressed using profanity	"These [IDENTITY] should shut the f*ck up once and for all."	hateful	160	98.8	**100.0**	**100.0**
	F11: Non-hateful use of profanity	I wish someone could get me the f*ck out of here."	non-hate	100	**100.0**	**100.0**	**100.0**
Pronoun references	**F12**: Hate expressed through reference in subsequent clauses	"I have met many [IDENTITY] and I hate every single one of them."	hateful	160	**98.8**	98.1	96.9
	F13: Hate expressed through reference in subsequent sentences	"Of course they are stupid. They are just [IDENTITY] after all."	hateful	152	98.0	97.4	**99.3**
Negation	**F14**: Hate expressed using negated positive statement	"[IDENTITY] lives do not matter."	hateful	160	97.5	**99.4**	**99.4**
	F15: Non-hate expressed using negated hateful statement	"No [IDENTITY] deserves to die." We shouldn't let [IDENTITY] suffer."	non-hate	152	*21.1*	*43.4*	**57.9**
Phrasing	**F16**: Hate phrased as a question	"Why is it that all [IDENTITY] are repulsive?"	non-hate	160	80.6	**99.4**	96.9
	F17: Hate phrased as an opinion	"I think that [IDENTITY] should all be killed."	hateful	152	98.0	**100.0**	99.3
Non-hate grp. ident.	**F18**: Neutral statements using protected group identifiers	We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hate	144	75.0	**77.1**	75.7
	F19: Positive statements using protected group identifiers	"I love [IDENTITY]." "[IDENTITY] are great."	non-hate	216	70.8	79.2	**87.0**
Counter Speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	190	*41.6*	54.7	**79.5**
	F21: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	156	62.8	51.9	**73.1**
Abuse against non-prot. targets	**F22**: Abuse targeted at objects	"I really can't stand cauliflower." "All cocktails like these are vile"	non-hate	65	**100.0**	98.5	98.5
	F23: Abuse targeted at individuals (not as member of a prot. group)	"You make me sick." "I will hurt you."	non-hate	65	**100.0**	**100.0**	**100.0**
	F24: Abuse targeted at non-protected groups (e.g. professions)	"Artists are parasites to our society." "Death to all capitalists."	non-hate	62	**98.4**	95.2	96.8
Spelling variations	**F25**: Swaps of adjacent characters	"I haet [IDENTITY]."	hateful	152	86.2	**90.8**	90.1
	F26: Missing characters	"[IDENTITY] are a cancr."	hateful	160	60.6	71.9	**83.1**
	F27: Missing word boundaries	"Deathto all [IDENTITY]."	hateful	156	80.8	81.4	**94.9**
	F28: Added spaces between chars	"I want to k i l l all [IDENTITY]."	hateful	190	73.2	**84.2**	**84.2**
	F29: Leet speak spellings	[IDENTITY] lives aren't worth sh1t."	hateful	190	**93.7**	**93.7**	92.1

B: $BERT_{base}$, H: HateBERT, R: RoBERTa

Table 31: Model accuracy (%) on HATECHECK functional tests with added *refugee* test cases. Highest scores for each test case are marked bold. The table is inspired by Röttger et al. (2020).

Overall, when considering the three types of evaluations in HATECHECK, it can be argued that RoBERTa is the most suitable model since it exhibits the least number of weaknesses in multiple tests. To alleviate the weaknesses associated with the models, more of these special cases that were difficult to handle for the models need to be added to the training set. This should increase the models ability to handle these cases as they are exposed to more instances than before.

5.4 REFUGEE CRISIS ANALYSIS

After having validated the models using multiple test datasets, it is now relevant to apply the best performing model, RoBERTa, to multiple unlabeled datasets, which consist of tweets surrounding specific international refugee crises for a constrained period. These datasets are obtained from Wolters and Olšavský (2021), and an explanation of the datasets are provided in section 4.3.2.3. By using these unlabeled datasets, it is possible to facilitate an analysis that focuses on the development of the amount of hate, offensive, and normal speech over time for specific crises. In Table 32, an overview of different types of statistics for each of the refugee crisis datasets is available.

Refugee Crisis	Days	n tweets	Hate speech (%)		
			Min	Mean	Max
Afghanistan	37	283,643	6.9	18.6	33.0
Channel Crossings	42	173,758	33.6	38.9	50.0
Greece-Turkey	42	137,462	15.4	23.8	32.8
Rohingya	61	29,432	1.9	7.1	16.6
Tigray	106	42,853	3.8	7.9	15.0

Table 32: Summary statistics for international refugee crises. The min, mean, and max values are the proportional amount of hate speech in % for each crisis.

From the table, it can be observed that the refugee crisis associated with the United Kingdom channel crossings contained the highest proportion of hate speech as it was 38.9% on average. Besides that, the proportion of hate speech on a single day reached 50%, which is the highest proportion among all the refugee crises. Furthermore, it can also be observed that the lowest proportion of hate speech for the United Kingdom channel crossings crisis is higher than all the other refugee crises' maximum proportion on a single day. Therefore, it can be argued that there is a lot of hate speech associated with this specific refugee

crisis compared to the other ones. Besides focusing on the refugee crisis that contains the highest proportion of hate speech, the table also shows that both the refugee crises related to Rohingya and Tigray experience the smallest proportion of hate speech with an average value of 7.1% and 7.9% respectively. Two of the refugee crises from Table 32 are analyzed more thoroughly in the subsequent sections. The figures and tables that are included in this analysis are available for the other three refugee crises in section A.3 in the Appendix.

5.4.1 Afghanistan Refugee Crisis

Figure 17 shows the volume of English tweets that are classified as hate, offensive, and normal speech per day for the Afghanistan crisis. Notable is the low total volume of tweets in the days before the fall of Kabul (trigger event), where 3,226 and 958 were classified as normal and hate speech respectively on the 14th of August 2021. The number of offensive tweets is negligible with between 0.5% and 2.5% of the total tweet volume. As result of the increasing press coverage of the event, the total volume of tweets rose fast until reaching its maximum on the 17th August 2021 with 25,066 normal and 6,912 hate speech tweets. After that, the total volume decreased quickly with a small rise especially for tweets classified as normal around the 24th to 26th August 2021.

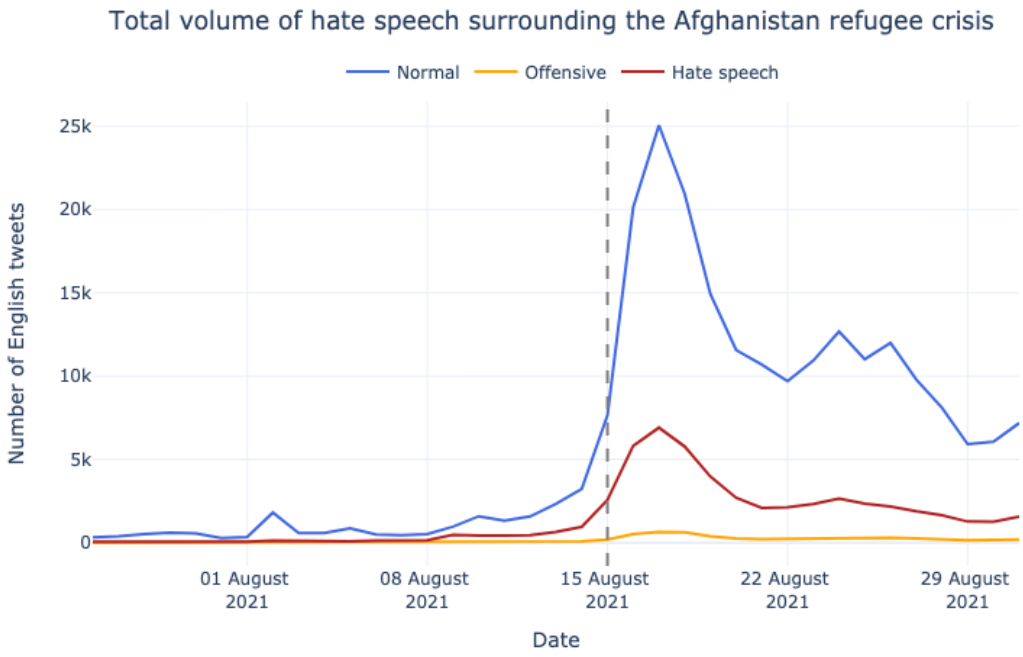

Figure 17: Development of the volume of hate, offensive, and normal speech in the period from 26-07-2021 to 31-08-2021 for the Afghanistan refugee crisis predicted with RoBERTa. The dashed line indicates the date of the trigger event.

Figure 18 shows the relative amount of hate speech compared to the total volume of tweets for a given day. Throughout the whole period, the proportion of hate speech ranges between 6.9% and 32.9%. From the figure, it can be observed that a high fluctuation appears in the days leading up to the trigger event, where the total volume of tweets is limited. Furthermore, this figure also includes a mean trendline, which is based on a rolling average of the hate speech proportion. By following that line, it can be observed that at the beginning of the period, the mean was 17.9% before it reached its minimum of 14.9% on the 4th August 2021. Afterwards, the mean trendline rose to 19.1% on the trigger event date before reaching its maximum plateau of 19.4% in the following days. On the 20th August 2021 the daily proportional volume falls below the mean trendline, signifying a slight downwards trend, with the rolling average ending up at 18.6%.

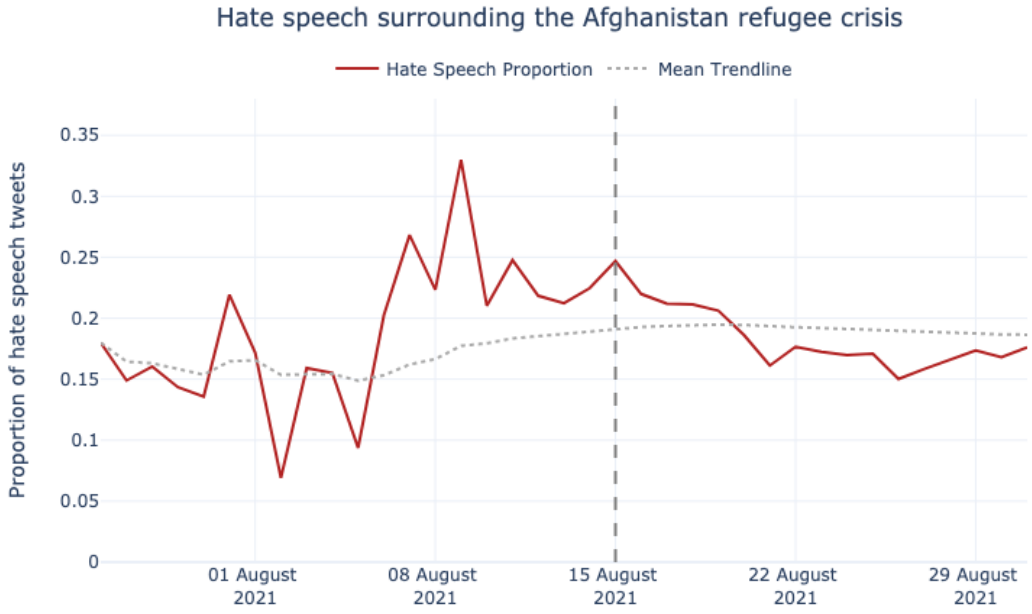

Figure 18: Development of the proportion of hate speech in % in the period from 26-07-2021 to 31-08-2021 for the Afghanistan refugee crisis. The vertical dashed line indicates the date of the trigger event.

FEATURE IMPORTANCE To provide a more detailed analysis of the selected refugee crisis, the importance of the features will also be calculated to identify the most influential features for this specific refugee crisis dataset. These features can then be compared with the most important ones for the other crises to see whether there a major differences. The feature importance is calculated by using the mean SHAP values, which is introduced in

section 4.6. Furthermore, it is important to state that the importance score is highly influenced by the context a specific feature appears in. The reason for this can be attributed to the fact that RoBERTa considers the context of a word when calculating its representation. To increase generalizability, only the features that appeared at least 2 times are displayed. In Table 33, the 10 most important features are presented together with the number of times they appear in the dataset.

Token	n	SHAP(Hate speech)
immigrants·	4	0.23
refugees·	110	0.17
refugees	36	0.16
refugee	5	0.15
migrants·	4	0.11
lim	2	0.11
refugee·	23	0.10
seekers·	2	0.10
lim·	5	0.09
asylum·	7	0.08

Table 33: Most important features for the hate speech class using the Afghanistan crisis dataset. The tokens in the list have occurred at least 1 time and the SHAP values are calculated using the mean value of all occurrences of a token in the dataset. Whitespace after a token is marked with a middle dot (·) indicating that the token is followed by another word. The token mapping is available in Table 36 in the Appendix.

5.4.2 Greece-Turkey Refugee Crisis

Figure 26 shows the volumes of hate, offensive, and normal speech for the Greece-Turkey crisis over time. From the figure, it can be observed that the discussion on Twitter surrounding the Greece-Turkey refugee crisis started to rise on the 27th of February 2020. Around this date, the Turkish President shared his intentions to open the Greece-Turkey border and let refugees into Europe. Therefore, it is considered the trigger event, which is visualized with a vertical dashed line in the figure. After his intentions were published, it triggered multiple discussions on Twitter, which can be observed in the figure as the volume of tweets continued to rise after this date. The public announcement of his decision to open up the border was made on the 29th of February 2020 (Stevis-Gridneff and Gall, 2020), which further increased the level of activity regarding this refugee crisis on Twitter. Here, it reached its highest on the 2th of March 2020, where 4,692 tweets were classified

as hate, 260 as offensive, and 10,277 as normal speech. In the subsequent period, both the volume of hate and normal speech started to return to their normal activity level again after having experienced a drastic increase over 4 days. Besides that, it is also noticeable that the volume of offensive speech stays consistent on a low scale throughout the whole time period.

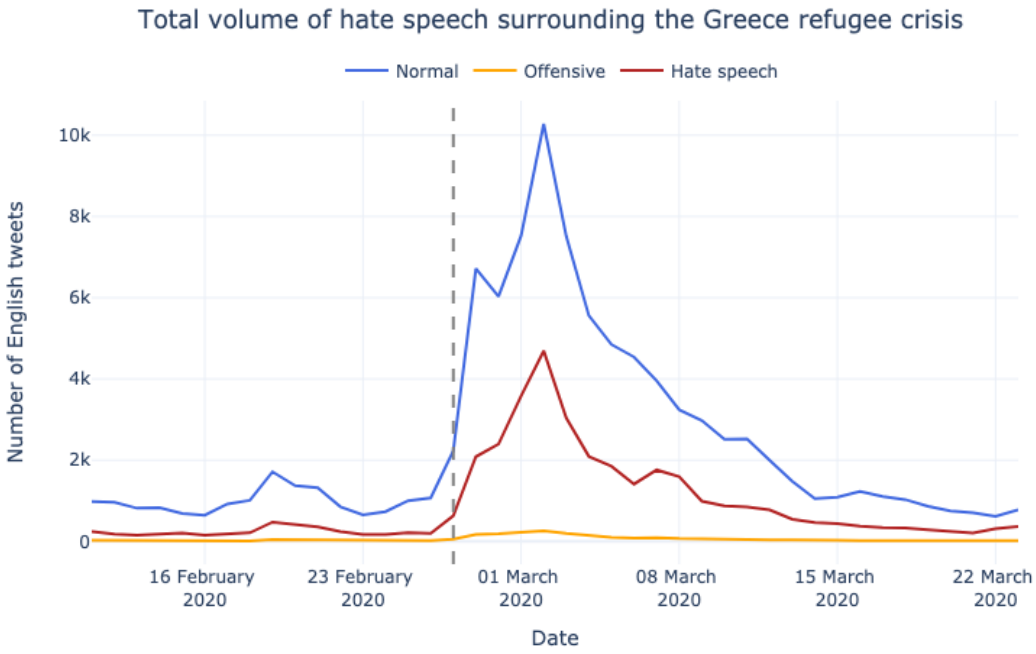

Figure 19: Development of the volume of hate, offensive, and normal speech in the period from 11-02-2020 to 23-03-2020 for the Greece-Turkey refugee crisis predicted with RoBERTa. The dashed line indicates the date of the trigger event.

Besides solely analyzing how the volume changes, it is also interesting to focus on how the relative amount of hate speech compared to the total volume of tweets acts over time. Figure 27 shows the proportion of hate speech for each day in the specific time period together with a trendline that is based on a rolling average of the hate speech proportion. From the figure, it can be observed that the proportion of hate speech fluctuates during the whole period. Also, before reaching the trigger event, which is represented by the dashed vertical line, the proportion of hate speech decreases to its lowest value of 15.5% in the whole period. After the trigger event on the 27th of February 2020, the data exhibits an increasing trend, where the proportion of hate speech is increased by a factor of 2 as it changes from 15.5% to 31.6% in only 4 days. Afterwards, it continues to stay at this high level of hate speech, which is also identified by the trendline. Besides that, the

highest proportion, 32.8%, is reached on the 22th of March 2020. To summarize, it can be argued that the date where the Turkish President shared his intention about opening up the Greece-Turkey border for Syrian refugees (trigger event) contributed to the increasing trend in the proportion of hate speech. This was further accelerated when the President publicly announced that his intentions were going to be implemented.

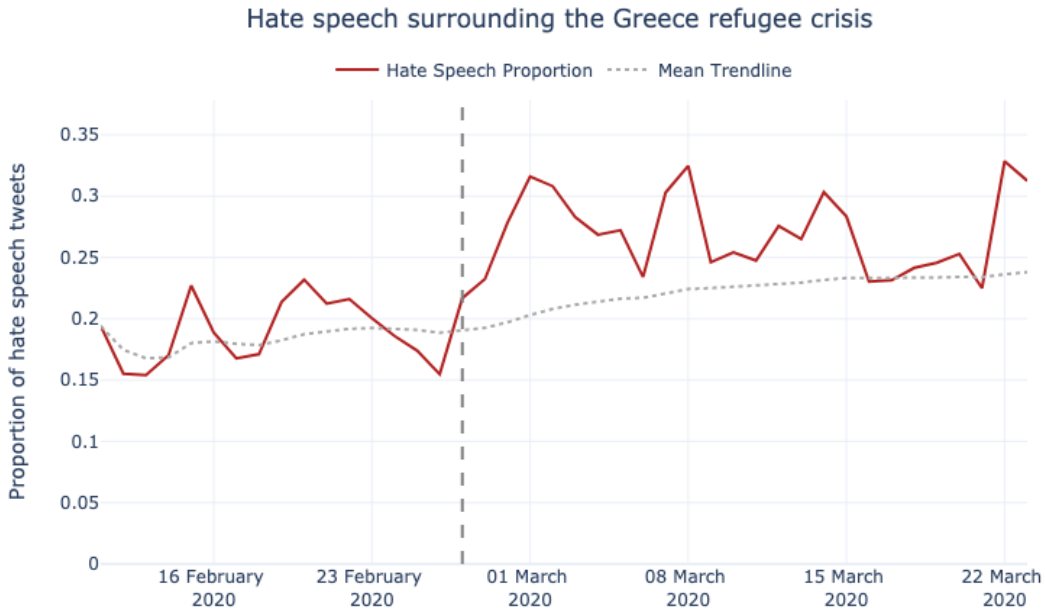

Figure 20: Development of the proportion of hate speech in % in the period from 11-02-2020 to 23-03-2020 for the Greece-Turkey refugee crisis. The vertical dashed line indicates the date of the trigger event.

FEATURE IMPORTANCE To be able to compare the most important features from the Greece-Turkey refugee crisis with the ones from the Afghanistan crisis, the importance of each feature in the dataset will also be calculated here. In Table 34, the 10 most important features are presented together with the number of times it appears in the dataset. The original word that refers to the "*uge*" token is available in Table 36 in the Appendix.

Token	n	SHAP(Hate speech)
migrants	7	0.29
immigrants·	21	0.21
refugees	24	0.17
immigrants	3	0.17
migrants·	35	0.17
refugee	6	0.16
uge	36	0.14
seekers·	2	0.14
migrant·	10	0.13
refugees·	67	0.13

Table 34: Most important features for the hate speech class using the Greece-Turkey crisis dataset. The tokens in the list have occurred at least 1 time and the SHAP values are calculated using the mean value of all occurrences of a token in the dataset. Whitespace after a token is marked with a middle dot (·), indicating that the token is followed by another word.

6 DISCUSSION

In the fifth step of the CRISP-DM framework, the outcomes of the models are reviewed and validated to see if the objectives specified are reached. This section, therefore, commences by answering the research question based on the findings from previous chapters. Besides that, the findings will be used to provide recommendations for UNHCR. Additionally, the organizational and societal implications will be discussed together with the limitations of this thesis. Lastly, further directions for the thesis are suggested.

6.1 ANSWERING THE RESEARCH QUESTION

The first section of the discussion addresses the research and supportive questions stated originally in section 1.1.2 of the thesis. More specifically, the supportive questions are discussed first as they are used to help answering the overall research question. To obtain a clear structure of the discussion, the questions are briefly repeated before being answered. As chapter 5 already contains detailed analysis of the results, the goal of this part is to connect the key findings to answer the research questions.

> *Q1: How well can deep learning models help to measure hate and offensive speech in general and for text surrounding refugees?*

The goal of this first question is to test multiple BERT models' ability to both detect hate speech in general and against refugees to see how well it can be measured and detected. From the results of the in-dataset experiment, it can be observed that by using a RoBERTa model, the highest macro F1-score of 81.0% is achieved compared to the scores obtained using a DistilBERT, BERT, and HateBERT model. Furthermore, when considering the per-class performance, RoBERTa also outperforms the other models on the hate speech and offensive class as it achieved a macro F1-score of 81.9% and 74.2% respectively. However, the baseline model (BERT) performs best for the normal class with a score that is 0.1 percent-

age point higher than the RoBERTa model. Since the difference is not significant enough, it can not be concluded that this model outperforms RoBERTa.

When looking at the classes in detail, the misclassifications from Figure 13 are of particular interest. These suggest that there are clear decision boundaries between the offensive and hate speech class, and that the distinction against the normal class is more difficult in the given train and test set. Only a hypotheses about a possible relationship between performance and available training data per class can be made here. The existence of a clear decision border can be confirmed when inspecting the feature importance for the offensive and hate speech class from Figure 14. The cumulative contribution of the top 100 most important tokens influencing the offensive class in turn have an almost equal negative impact on the normal class and their effect on the hate speech class is close to neutral. Consequently, the top 100 tokens with the highest contribution towards the hate speech class only have a slight negative effect on the offensive class, but decrease the likelihood of the model output being normal. These observed decision boundaries provide evidence that the model is able to separate well between the offensive and hate speech classes.

Even though the different models perform similar to each other in the in-dataset experiment, there are evident performance differences when comparing the models on data that contains refugee-related data provided by UNHCR. When consulting the results, the RoBERTa model performs best with an accuracy of 73.5% outperforming the second best model by a margin of 2.1 percentage points. Based on both the in-dataset evaluation and the evaluation on the refugee-related dataset, it indicates that it is more beneficial to use a model that is both pre-trained for a longer time and pre-trained using much more data than the other models (Liu et al., 2019).

On the other hand, it is also important to consider the cost associated with fine-tuning a larger model. In this case, it can be discussed whether it is worth to use a model that achieves an accuracy that is 2.1 percentage points higher than both the baseline model and HateBERT but requires a longer time to fine-tune. The trade-off between fine-tuning cost and accuracy is an important discussion, but in this thesis, the main goal is to achieve the

best possible model for classifying hate speech, which is why the increased fine-tuning time does not have a big influence. Besides comparing the performance of all the models, the results also indicate that it is possible to detect hate speech surrounding refugees by fine-tuning the models using 12 different hate speech datasets.

The models performance on refugee specific data is additionally confirmed by the HATECHECK functional test set, where the RoBERTa model shows very high accuracy on both the "*Refugees*" and the discursively related "*Immigrants*" target groups, with accuracy scores both higher than 91.0%. This indicates that the particular functional weaknesses tested by HATECHECK and exhibited by the model are less severe for the refugee related groups than for the other target groups. These weaknesses are mostly featured in difficult tests of non-hate. This mainly includes negation (**F15**), non-hateful statements against protected group identifiers (**F18** and **F19**) as well as counter speech (**F20** and **F21**).

Overall, the additionally pre-trained HateBERT model performed better than the baseline BERT model. For most of the hateful functional tests, HateBERT was able to achieve slightly better results than RoBERTa. This especially become evident for samples that consisted of hateful slurs (**F7**). However, the RoBERTa model completely outperformed HateBERT on the majority of the non-hateful tests. Nevertheless, the best performing model over all metrics is clearly RoBERTa, outperforming all other tested models, most decidedly in the HATECHECK test suite.

Answering *Q1*, one can conclude that deep learning models like RoBERTa can indeed help to detect and measure offensive and hate speech in general as well as hate speech against refugees. Offensive speech against refugees was neither observed to be common or tested against. Conjointly, the results were confirmed to be satisfactory for production use by UNHCR in bi-weekly result meetings.

Q2: Which features are important for identifying hate and offensive speech?

This question's aim is to reflect on the feature importance sections and connect the core findings of the individual datasets with each other. Overall, the feature importance between all the models has been observed to be very similar. However, there are interesting differences when analyzing the feature importance for the best performing model, RoBERTa, on different datasets and classes, which will be examined to answer this question.

When examining the offensive class for the test dataset that consists of data from the same domain as the training dataset, as presented in Table 26, it can be observed, that the token with the highest feature importance score is "*stupidity·*", which has a SHAP value of 0.55. This is 0.15 points higher than the top hate speech token, "*igger·*". Moreover, the top 14 offensive tokens all have a higher SHAP value than the top hate speech tokens from the same test set. This indicates that the offensive tokens in the test dataset are more influential than the hate speech tokens. Overall, the tokens increasing the likelihood of a sample being classified as offensive are "*stupid(ity)*", "*pussy*", "*ridiculous*", and "*bitch·*". These tokens all have a negligible negative effect on the hate speech class.

Some of the top offensive tokens are particularly interesting, like the token "*ches·*". It is one of the few tokens that has a positive impact on the hate speech class with a mean SHAP value of 0.10, while being one of the most important tokens for the offensive class. The token is part of many different words, with only some being offensive or hateful. From the dataset, it can be observed that the token probably comes from one of the words "*bitches*", "*churches*" or "*turkroaches*". Since the token represents multiple words, it makes it hard to generalize the findings because it is difficult to figure out which of the original words are the most influential one. Based on the importance scores, it can be argued that the use of individual tokens often strongly support classifying a sample as offensive in this dataset. This can be attributed to a reduced need for context when classifying offensive speech, for humans, as well for the model. This results in higher mean feature importance scores for top tokens.

The feature importance for hate speech tokens is even harder to evaluate because the con-

text of the samples is much more important. As a result, the mean feature importance for each feature is much lower than for the offensive tokens. Besides that, when analysing the results from Table 26, the 14th most important token, "*immigrant·*", only has half the SHAP value when compared with the most important the token "*igger*". Notable is that when this slur is used in singular form, as in "*igger·*" the SHAP value decreases by 0.13, which marks the biggest difference between same word stems. From the result, it is also interesting that the other 3 top tokens are always followed by other characters or punctuation.

When comparing with all the refugee related datasets from Figure 15, 33, and 34, it can be observed that tokens related to "*immigrants*", "*refugees*", "*refugee*", and "*migrants*" dominate the hate speech feature importance with SHAP values around 0.20. This is in line with UNHCRs findings as they have previously identified that if these types of words appear in a sequence, it is most often associated with hateful content. Other recurring tokens with approximately half the SHAP value are "*asylum·*", as well as variations on "Islam" and "Muslim", which is represented by the tokens "*lam*" and "*lim*".

Important features such as "*rapist*", "*criminals*", and "*terrorist*" are only present in the limited UNHCR test dataset. Besides that, a difference of 0.06 in the importance score between the top token "*refugees*", normally followed by punctuation like in "*refugees!!!*" and the free-standing "*refugees·*" can be found. Even though punctuation like the exclamation mark have negligible SHAP values for the hate speech class, from this it can be inferred that certain words followed by them increase the likelihood of a sample being hate speech. This is similar for the dataset surrounding the Greece-Turkey crisis but not for the Afghanistan crisis.

Additionally, differences between the two refugee crises can be identified. For example, the most hateful token in the Greece-Turkey crisis "*migrants*" followed by punctuation is not present in the Afghanistan crisis, hinting that these crises have different focuses for authors of hate speech. In some cases, the same token can have a negative impact on the likelihood of a sample being hate speech for one crisis, like with "*illegal·*" token, which is with -0.03 slightly negative in context of Afghanistan, but contributes to hate speech with 0.06 in the Greece-Turkey crisis.

To summarize *Q2*, it can be stated that SHAP values can be used to approximate the mean feature importance of tokens for a dataset. More specifically, tokens that are important for the offensive class are much more influential than the hate speech related tokens. Besides that, when looking at the hateful text targeted against refugees, it can be extracted that tokens surrounding refugees and migrants are used interchangeably, and that they have an overall negative connotation in these datasets. Yet, the actual classification is highly dependent on the context of each sample. Furthermore, it is also important to state that these importance scores give an indication of their influence in the specific datasets that are used in this thesis. Therefore, if the tokens appear in other contexts, the influence might be different because the BERT models represents words based on their context.

> *Q3: How does hate and offensive speech change over time in the context of specific international refugee crises?*

To answer this question, data for the following five refugee crises; Afghanistan, United Kingdom channel crossings, Greece-Turkey, Rohingya, and Tigray, was obtained from Wolters and Olšavský (2021). From the overall statistics of the dataset, it was identified that the United Kingdom channel crossings contained the most amount of hate speech as the data showed an average of 38.9% hate speech per day throughout the period that was analyzed. On the other hand, the refugee crises associated with Rohingya and Tigray contained the least amount of hate speech. One potential reason for this could be that these two crises are not widely covered internationally compared to the other ones. Instead, there might be more Twitter activity on a local level, and it would therefore make sense if more hate speech was present there. When interpreting these results, it is important to state that this only represents a snapshot in time and includes a subset of tweets from Twitter. That being said, the main point with answering this question is to prove how trends and patterns can be derived from such type of analysis.

A deeper analysis for Afghanistan and Greece-Turkey was performed as these two refugee crises exhibited the most interesting findings. The data that was used to analyze the Afghanistan refugee crisis was obtained from the period when Kabul was overtaken (trig-

ger event) by Taliban in August 2021. From the analysis, it was identified that both the total number of tweets and hate speech started to increase a lot around the trigger event, which was the 15th of August 2021. It continued to rise in a 2 days period, where it reached the maximum amount of Twitter activity the 17th of August 2021. Afterwards, both the amount of hate, offensive, and normal speech started to decrease again. Furthermore, the development of the proportion of hate speech over time was analyzed together with a rolling average trendline. The resulting visualization showed that the proportion of hate speech was fluctuating a lot before reaching the trigger event. The reason for this can be attributed to the fact that the number of total tweets in the period before the trigger event was limited. Besides that, it was observed that only after a couple of days after the trigger event, the proportion of hate speech decreased below the mean trendline. This suggests that the trend is starting to experience a decrease. By following the behaviour of the trendline, it can be argued that it is somewhat stable with a minor increase throughout the whole period even though a large increase appeared.

The data that was used to analyze the development of hate, offensive, and normal speech over time for the Greece-Turkey refugee crisis was obtained from the period where the Turkish President shared his intentions about opening up the Greek-Turkish border such that refugees from Syria could enter Europe (trigger event). This was met with resistance from the Greek side. From the analysis, it was observed that this event triggered multiple discussions on Twitter, especially when the President's intentions was published on the 27th of February 2020. The consequences of this was that the Twitter activity, more specifically the amount of hate and normal speech, surrounding this refugee crises exploded in a period of 4 days. Afterwards, it started to return to the normal level of activity again. When considering the analysis of hate speech proportion in relation to the total number of tweets, it was found that the proportion of hate speech fluctuated throughout the whole time period. What is interesting is that the hate speech proportion increased by a factor of 2 in a matter of days after the trigger event, since it went from 15.5% to 31.6%. Also, when focusing on the mean trendline, it was found that the overall trend was exhibiting an increasing trend after the trigger event because the proportion of hate speech remained high, even though the total number of tweets decreased 4 days after the trigger event.

When comparing the findings for the two refugee crises, it was observed that there was not any trends or patterns associated with the text that was classified as offensive. It was also observed that both refugee crises experienced a drastic increase in both the amount of hate and normal speech for a couple of days before it started to return back to the usual level again. This could possibly be attributed to the fact that these datasets are centered around a trigger event, and that it got a lot of international media attention. This was not the case for the Rohingya and Tigray refugee crises, which are analyzed in section A.3.

Based on these findings, it is relevant to discuss what this can be used for in UNHCR's setting. By performing this type of analysis, makes it possible to track the Twitter activity regarding specific refugee crises over time to see how the discussions evolve. More specifically, trends and patterns can be derived, which can provide an indication of when UNHCR needs to implement preventative measures to try to control the situation. Therefore, it can be used in the context of a warning system. The findings of the analysis showed that it takes a couple of days before the amount of hate speech reaches its highest, therefore, UNHCR has the opportunity to affect the discussions before it escalates even further. To summarize, it is important to mention that these results cannot be easily generalized because they only represent a subset of all the available tweets in a snapshot of the time surrounding these refugee crises.

> *RQ: How can hate speech against refugees on social media platforms be detected and measured using Natural Language Processing methods?*

Based on the discussions of the supportive questions, it can be argued that hate speech against refugees on social media platforms can be detected and measured using BERT models that build on transformer-based architectures, which are able to leverage large amounts of textual social media data to confidently detect hate speech in the context of refugees (*Q1*). These models makes it possible to represent the meaning of words by including the context that they appear in. This increases the contextual understanding of the models, and makes it easier for the them to correctly predict hate and offensive speech. Even though offensive speech is not part of this overall research question, it is added as an additional

feature of the model such that it is possible for UNHCR to also measure the development of offensive speech on social media targeted at refugees. As a result, it is part of the supportive questions.

Extracting the feature importance makes it possible to measure the impact individual tokens have on the output of a model in the context of an international refugee crisis ($Q2$), which enables further analysis for counter measures. The model predicts hate and offensive speech sufficiently well, which allows it to be used to measure how the amount of hate and offensive speech changes over time for different international refugee crises. By applying the model on unlabeled data surrounding these crises, it can be viewed as a proof-of-concept for UNHCR to use this model productively in their activities to analyze the full extend of relevant refugee crises ($Q3$).

6.2 RECOMMENDATIONS

Based on the findings in this thesis, it is possible to provide recommendations that can be implemented by UNHCR such that the organization is able to take advantage of the work from the thesis in a way that it generates value and improves internal processes. The recommendations that are suggested here are based on actionable insights from the findings of this work.

6.2.1 Root Cause Analysis

This thesis only provided an analysis of two specific refugee crises in a short time frame using a subset of tweets. Therefore, it is recommended to perform a deeper analysis of the individual refugee crises to better understand root causes and central actors associated with hate speech published on Twitter over time. Besides that, it also becomes easier to identify when the amount of Twitter activity surrounding each crisis exceeds the regular level. To reduce the amount of manual work, the classification model that was developed in this thesis should be integrated into UNHCR's internal systems. This also makes it possible to analyze a much higher proportion of tweets than before. To obtain a more insightful analysis, the recommendations introduced below can be followed.

- **CREATE A REPOSITORY OF REFUGEE CRISES DATA** To obtain a deeper understanding of each refugee crisis, it is recommended to collect more comprehensive datasets than what was used in this work from more sources than just Twitter. Additionally, the data should cover a longer time period, if possible multiple years, such that long term trends can be understood. Furthermore, since the classification model was applied to unlabeled data for the refugee crises analyzed in this thesis, a subset of tweets should be annotated such that the performance of the model can be evaluated for each specific crisis as well.

- **COMBINE HATE SPEECH PREDICTIONS WITH OTHER META-DATA** To better identify the root causes of hate speech surrounding refugees, it is recommended to enrich the analysis with meta-data as this can give a more comprehensive understanding of the specific crises. By implementing this recommendation, UNHCR is going to contribute to the "*Commitment 2: Addressing root causes, drivers and actors of hate speech*" that is specified in the strategy and plan of action report regarding hate speech that was issued by the United Nations (United Nations, 2020).

6.2.2 Hate Speech Monitoring

To improve UNHCR's internal processes, it is also recommended to use the classification model to continuously monitor the development of hate, offensive, and normal speech surrounding international refugee crises. More specifically, the recommendations outlined below can be implemented.

- **IMPLEMENT MONITORING SYSTEM** Based on the findings of this thesis, it is recommended that UNHCR should implement an internal monitoring system that makes it possible to track how the amount of hate, offensive, and normal speech evolves on a day to day basis for particular areas. More specifically, the analysis of the predictions should be presented in a dashboard. Besides that, it can also be used as an early warning system because it is possible to observe when the proportion of hate and offensive speech starts to rise for a specific crisis. By implementing this recommendation, UNHCR is going to contribute to the "*Commitment 1: Monitoring and analysing hate speech*" of the UN's strategy and plan of action on (United Nations, 2020).

- **CONTINUOUSLY RETRAIN CLASSIFICATION MODEL** To constantly be able to provide a reliable monitoring, it is also recommended to retrain the classification model occasionally with new data. This is crucial because hate speech is a complex phenomenon and the understanding of what is considered as hate speech is dynamically changing over time. To implement this recommendation, it is required that the annotations of the training dataset is reviewed or new annotated text data is provided to the model.

6.2.3 Contribute to the Hate Speech Research Field

According to the literature, work surrounding hate speech detection against refugees is limited and almost not existing. Therefore, it is recommended that UNHCR should have a higher focus on contributing to the research field to increase the awareness of this area. When implementing this recommendation, the bullet points below provide some recommended action points that are required for this to work efficiently.

- **CREATE HIGH QUALITY HATE SPEECH DATASETS SURROUNDING REFUGEES** The first action point is to create datasets that consist of high quality annotations because it was observed both in this thesis and the literature that the data quality is crucial when training a hate speech detection model to obtain accurate predictions. Also, UNHCR is considered the domain expert in the area of hate speech targeting refugees, which means that they have a solid foundation for providing reliable annotations. Furthermore, since hate speech is a term that changes dynamically over time, it is important that UNHCR reviews the annotations of the datasets periodically because due to changing societal norms, what is not considered as hate speech today, might be considered hateful in the future.

- **USE BEST PRACTICES FOR ANNOTATING** Based on the literature, multiple best practices have been collected, which will be recommended here. First of all, the most important thing is that each tweet needs to be annotated by multiple people because it increases the reliability of the annotations in the end. Besides that, the annotators need to be well-educated for the task at hand since hate speech is a complex phenomenon where multiple definitions and understandings of the term exist. To educate annotators properly, thorough guidelines need to be distributed. More specifically, these guidelines should provide definitions and examples of hate and offensive speech. An

explanation of how difficult cases need to be treated should also be included in the guidelines. Furthermore, annotators should practise on a benchmark set such that their performance can be evaluated. This is both beneficial for UNHCR and the annotators. To evaluate the agreement among multiple annotators, it is important to use an agreement metric such as Fleiss' Kappa (Fleiss, 1971).

- **SHARE ANNOTATED DATASETS WITH ACADEMIA** After having created high quality datasets with hate speech surrounding refugees, it is recommended that these datasets are shared with academia. The intention with publishing these datasets is to stimulate more research within this area, which will benefit UNHCR in the end because multiple experiments are going to be carried out. As a result, it will be possible to take advantage of potential suggestions for model improvements.

- **SET UP A SHARED TASK TOGETHER WITH ACADEMIA** Besides sharing hate speech datasets surrounding refugees, it is also recommended that UNHCR should set up a shared task about hate speech detection targeted at refugees in academia together with some researchers. The benefit of this is that the awareness of the problem with hate speech against refugees is increased. As a result, multiple research teams will possibly participate in the shared task, and the findings of their research can be used to improve UNHCR's classification model in the end.

To summarize, it can be argued that contributing to the research field is a cost-effective way to both validate the datasets and obtain ideas for further improving the internal hate speech detection system. By increasing the awareness of the problem with hate speech targeted at refugees in academia, it will possibly increase the amount of available research within this area, which is beneficial for UNHCR.

6.3 IMPLICATIONS

This section discusses the different implications that the findings of this thesis have on both an organizational and societal level.

6.3.1 Organizational Level - UNHCR

The United Nations (UN) has issued a strategy and plan of action of how to tackle hate speech against the humanity (United Nations, 2020), where UNHCR is considered the lead-

ing agency until 2024. So far, UNHCR has mostly relied on manual processes to investigate the content of tweets that are posted online. More specifically, this includes identifying risky areas, root causes, actors, and frequent phrases that are being used in a hateful context to create an overall picture of the situations. The findings of this thesis show that hate speech surrounding refugees can be detected using state-of-the-art deep learning methods. As a result, it is possible to complement the work that UNHCR already carries out. The results and methods from the thesis were validated by UNHCR, who was very satisfied with the accuracy of the RoBERTa model.

When considering the implications for UNHCR, it can be argued that it is possible to automatize parts of the manual processes. More specifically, the best performing model can be plugged into the internal systems to perform regular scans, which can help with identifying potential risky areas. By taking advantage of the work that has been carried out in this thesis, entails a variety of benefits. For UNHCR, it will be much faster and easier to flag hate speech targeted at refugees on social media platforms because right now UNHCR does not have enough manual resources to go through all tweets themselves. This would make it possible to limit the impact of online hate speech as preventative measures can be taken much faster than before. From the identified hateful content, the risky areas, root causes, actors, and influential words/phrases can be identified faster as well, which increases the overall understanding of specific trends. This will in the end make it easier for the organization engage in online discussions such that the hateful content can be neutralized. This is crucial because online hate speech is highly associated with physical violence and oppression (Williams et al., 2020), and the United Nations aims at fostering a peaceful world.

In the thesis, it was also shown that the model makes it possible to track the development of hate, offensive, and normal speech over time on a social media platform. The benefit of this is that UNHCR can obtain a better and more detailed understanding of when hate speech in a certain area starts to rise together with how it evolves over time. In the strategy and plan of action report issued by the United Nations (United Nations, 2020), multiple action points were made available to the relevant agencies. As this thesis aims at

detecting hate speech surrounding refugees in collaboration with UNHCR, it contributes to the following three action points "*Commitment 1: Monitoring and analysing hate speech*", "*Commitment 2: Addressing root causes, drivers and actors of hate speech*", and "*Commitment 6: Using technology*" (United Nations, 2020).

6.3.2 Societal Level

When considering the societal implications, one has to do so from the assumption that UNHCR is using the methods and models presented in this work, and employs them in a monitoring system as described in previous chapters. This would enable the organization to detect changes in resentment against refugees, especially increases in the host societies' feelings of hostility, anger, and mistrust toward refugees, asylum seekers, and immigrants.

Furthermore, it will allow UNHCR to take action earlier, in the form of either advising and representing refugees towards governments or working with social media platforms to decrease the amount of hate speech on their platforms through various methods. Another way action can be taken is through increased media coverage of a crisis, as a result of UNHCR choosing to highlight the responsibility of a platform in regards to hate speech in the context of a specific refugee crisis. However, as it was observed in section 5.2 that this can result in an increase of the total amount of hate speech as the crisis makes headlines. Additionally, this could have the effect that shareholder value of the mentioned social media platform decreases as a result of negative media coverage. This increases the likelihood of inducing change in the platforms policies or algorithms in regards to hate speech. Moreover, it has to be mentioned that these platforms are in a dilemma, as increases in restrictions can lead users to leave the platform, conversely negatively impacting shareholder value. Additionally, the models that are developed in this thesis makes it possible to detect if preventative strategies effectively decreases the total amount of hate speech targeting refugees. The intention of all these measures is to support increasing sympathy towards refugees, while reducing violence against catalyzed by online hate speech with the goal of preventing future atrocity crimes.

6.4 LIMITATIONS

This section highlights the limitations associated with this thesis, which have an effect on the findings.

6.4.1 The Dynamic Nature of Hate Speech

Classifying hate speech is a complex task both for humans and deep learning models. The reason for this can be attributed to the fact that the understanding and interpretation of the term changes over time. This means that what is not considered hate speech today might be considered as hate speech in the future. Therefore, it can be argued that the nature of hate speech is dynamic due to the fact that both linguistic and social norms evolve over time. As a result, it is more considered as a conceptual limitation rather than a limitation related to the methodology and design of the thesis. The consequence of this is that it makes it even more difficult to classify text samples as hate speech because it requires frequent updates of the training data, where the annotations either are reviewed or new high quality annotated data is added. Also, these annotations need to be handled by really well-trained annotators because it is difficult to both formalize and define rules for what is considered as hate speech. In some scenarios, it also depends on whether the person/group that is being targeted interprets it as an insult. More specifically, in this thesis the changing nature of hate speech was observed when investigating the 12 datasets that were combined. Here, it became clear that there was not a general consensus to what hate speech is, which affects the models negatively.

6.4.2 Combining Multiple Datasets for Training

In this thesis, 12 different datasets have been combined together into one large dataset, which are used to train deep learning models to detect hate speech surrounding refugees. The motivation for choosing this approach was that the availability of high quality hate speech datasets was limited, therefore, the idea was to combine multiple of the available ones to test whether it was possible to take advantage of these. Even though, it was shown in the results that hate speech surrounding refugees could be detected by combining multiple datasets from different domains, the approach is still questionable. The reason for this can be attributed to the fact that different collection strategies and label definitions have been used by each research team to obtain the data that are used in the thesis. Besides that,

different annotation strategies have also been applied. This means that all the data has been annotated by people with multiple skill levels, and the guidelines that have been provided to the annotators are also different. Additionally, each annotated text sample has not been validated by the same number of annotators. Since hate speech detection is considered as a complex task, it is crucial that the data contains high quality annotations provided by expert annotators.

All this has an influence on the overall quality of the data because of the inconsistencies between each dataset. To avoid too many major inconsistencies, a thorough investigation was performed before picking out the datasets such that it was only the most recent ones and the ones with the highest quality (defined on many parameters) that were chosen.

6.4.3 Refugee-Related Evaluation Dataset

To validate whether the 12 different datasets from multiple domains could be used to detect hate speech surrounding refugees, a dataset with 98 hateful tweets targeting refugees was provided by UNHCR. Based on the results, it can be observed that the deep learning models were able to detect hate speech surrounding refugees. On the other hand, it can be questionable whether the number of tweets in the dataset was enough to perform a thorough evaluation. The reason for this is that the dataset only consists of a limited number of examples, which means that the variety is low, and therefore, it can be argued that it is not representative enough. By having more data to test the models in a refugee-related context, would possibly have affected the performance of the models as there would have been a larger variety of test cases. The authors of the thesis tried to obtain a larger test set, but UNHCR was not willing to provide more examples as they were satisfied with the result.

Furthermore, a subset of the tweets in the dataset was orignally written in a different language than English, and was therefore, translated into English by UNHCR using Google Translate. This constitutes a limitation because there is a high chance that context and language-specific features disappear in the translation process. Besides that, it can also be argued that there is a limitation associated to the annotation of the data as this was performed by a single employee from UNHCR. It has already been identified that hate speech detection is a complex task, and even though, the annotator was considered an

expert, using multiple expert annotators would possibly have increased the quality of the annotations. Here, it would be possible to calculate Fleiss' Kappa (Fleiss, 1971) to measure the aggreement among the annotators. The tweets that were misclassified by the RoBERTa model was further analyzed in the results, and it was observed that some of the misclassifications were a result of a human error. This could potentially have been mitigated using multiple expert annotators.

6.4.4 Feature Importance

When evaluating the performance of the different deep learning models, the feature importance was also calculated to measure the impact of each feature on the prediction. Since this thesis only uses state-of-the-art BERT models, which are built using the transformer architecture provided by Vaswani et al. (2017), each feature highly depends on the context that it appears in. Therefore, it is difficult to state which features are generally important for the hate speech class. Instead, it was only possible to show which features that were important when classifying hate speech in a specific dataset. As a result, the findings cannot be generalized due to the dependency of context.

6.5 FUTURE WORK

This section introduces multiple directions for further developing on the research that was conducted in this thesis. More specifically, the ideas for future work mainly focus on improving the data foundation and models.

6.5.1 Dataset Improvements

To improve the overall quality of the training data, the impact of each of the 12 datasets that are used for training the models should be investigated. In this thesis, the datasets were combined without any further experiments regarding the contribution of each dataset. Therefore, it might be the case that some of the datasets are either irrelevant or hurting the performance of the applied models. To obtain this type of insight, an ablation study should be performed as the behaviour of the models can be observed when for example removing one dataset at the time. This would both increase the understanding of the datasets, increase the performance of the models, and reduce the time needed to fine-tune the models if some of the datasets are removed.

6.5.2 Improve Contextual Understanding

Based on the discussions with UNHCR, and the falsely classified tweets from the refugee data (see section 5.2), it can be argued that the contextual understanding can further be improved using meta-data such as for example profile bios, profile pictures, network information, and pictures used in tweets etc. This would make it possible for the models to take this additional information into consideration when classifying whether a text sample is considered hateful or not. An example to illustrate this includes the use of the n-slur, which is socially accepted only when used in a reclaimed manner by a black person, but never for non-blacks. As a result, if the model is able to process profile pictures, it would be possible to extract information from these to get an understanding of the author of the text if it is available. Besides that, many posts also include images, which are often connected to the text in the posts. Again, it would be more suitable to have a model that could take these images into consideration as well as it will provide the models with more context, which possibly could improve the models' understanding of the case.

6.5.3 Model Improvements

The HATECHECK dataset from Röttger et al. (2020) was used to identify the areas where the models were struggling, and the results indicate that the models mainly fail at handling negations and counter speech in a non-hateful context. To further improve the performance and robustness of the models within these areas, the training dataset should be expanded with such text examples to become better at handling these cases. This could be implemented using multiple rounds as in Vidgen et al. (2020). More specifically, it means that in each round, new examples from a certain area will be added to the training set and the models will be evaluated to see whether it improves the performance of the models. This will continue for multiple rounds until a satisfied result is achieved.

7 CONCLUSION

This thesis has been conducted in collaboration with The United Nations High Commissioner for Refugees (UNHCR), and aimed at answering the research question "*How can hate speech against refugees on social media platforms be detected and measured using Natural Language Processing methods?*". Additionally, three sub-questions were defined to guide the answering of the research question. By investigating the academic literature, it was discovered that work and datasets regarding hate speech detection surrounding refugees were limited. Consequently, this thesis combined 12 different datasets, which have been used to detect hate and abusive speech in other contexts, while ensuring that the same understanding of what constitutes hate speech was maintained. To be able to combine the datasets into one large training dataset, a standardization of the labels were performed to group the labels into the following three classes; hate, offensive, and normal speech. The idea with this approach was to test whether it was possible to take advantage of hate speech datasets that have been used in different contexts to detect hate speech targeted at refugees.

To detect how well deep learning models are able to measure hate and offensive speech in general and in the context of refugees, three types of experiments were carried out. First, an in-dataset evaluation was performed, which was followed by an evaluation of the models in the context of refugees. To perform the second type of evaluation, a dataset with 98 hateful samples, which was provided by UNHCR, was used. In these experiments, a DistilBERT, $\text{BERT}_{\text{BASE}}$, HateBERT, and RoBERTa model were tested, and it can be concluded that the RoBERTa model performed best both in the in-dataset and refugee-related experiment with a macro F1-score of 81% and an accuracy of 73.5% respectively. The result on the refugee-related dataset were confirmed to be satisfactory for production use by UNHCR in bi-weekly result meetings. To outline the weaknesses of the RoBERTa model, the HateCheck dataset from Röttger et al. (2020) was applied. Based on the findings in the analysis,

it can be concluded that the model exhibited weaknesses regarding negations, non-hateful statements against protected group identifiers, and counter speech.

Besides testing different models, the most important features were extracted from the hate speech dataset provided by UNHCR. The results showed that the tokens "*refugees*", "*refugee*", "*igrants*", and "*immigrants*" appeared to be among the most influential features. This finding is in line with UNHCR as they have previously identified that the presence of these words is most often associated with hateful content. Furthermore, to analyze how hate and offensive speech changed over time for specific international refugee crises, data has been obtained from Wolters and Olšavský (2021). More specifically, from analyzing the Afghanistan and Greece-Turkey refugee crises, in a snapshot of the time, it was shown that the amount of hate speech increased significantly in a couple of days after a trigger event took place before it started to return back to the normal Twitter activity level again. The Greece-Turkey crisis stood out because the hate speech proportion trendline (rolling average) exhibited a minor increasing trend after the trigger event, even though the overall activity level was decreasing. Besides analyzing the change over time, the most important features were also extracted for each of these two refugee crises, and it can be concluded that the findings is similar to the ones from the UNHCR refugee dataset.

Based on the findings in this thesis, it can be concluded that hate speech targeted at refugees on social media platforms can be detected using a variety of BERT models, which possess the ability to understand context. More specifically, the results indicate that a more robust model (RoBERTa), which is pre-trained using a large amount of data and for a long time, is superior when considering all the different experiments. As a result, it is recommended that UNHCR should integrate the classification model into their internal systems as it will facilitate a better foundation for analysis, and it will also limit the manual resources needed to analyze tweets. This means that it will be possible to flag hateful tweets faster and easier.

Multiple works regarding hate and abusive language detection in different contexts already exist in the academic literature, but to the best of the authors knowledge, this work constitutes one of the initial works in the English language regarding hate speech detection

in the context of international refugees. In this thesis, multiple state-of-the-art deep learning methods were tested, which serve as baseline for further research. The importance of well-performing hate speech detection systems is stressed because online opinionated text can both result in a distorted image of refugees and leads to violence or atrocity crimes against refugees (Williams et al., 2020). To better mitigate the consequences for refugees, more research within this specific domain is required.

BIBLIOGRAPHY

Alammar, J. (2018). The Illustrated Transformer. Available online at: https://jalammar.github.io/illustrated-transformer/. Accessed on: 12/05/2022.

Alammar, J. (2019). A Visual Guide to Using BERT for the First Time. Available online at: http://jalammar.github.io/a-visual-guide-to-using-bert-for-the-first-time/. Accessed on: 04/04/2022.

Alshenqeeti, H. (2016). Are Emojis Creating a New or Old Visual Language for New Generations? A Socio-semiotic Study. *Advances in Language and Literary Studies, Australian International Academic Centre 7*(6), 56–69.

Arcila-Calderón, C., D. Blanco-Herrero, M. Frías-Vázquez, F. Seoane-Pérez, and S. Bratosin (2021). Refugees Welcome? Online Hate Speech and Sentiments in Twitter in Spain during the Reception of the Boat Aquarius. *Sustainability 13*(5).

Badjatiya, P., S. Gupta, M. Gupta, and V. Varma (2017). Deep learning for hate speech detection in tweets. In *26th International World Wide Web Conference 2017, WWW 2017 Companion*, pp. 759–760. International World Wide Web Conferences Steering Committee.

Bahdanau, D., K. Cho, and Y. Bengio (2014, 9). Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Barbieri, F., J. Camacho-Collados, L. Neves, and L. Espinosa-Anke (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on*

Semantic Evaluation, Stroudsburg, PA, USA, pp. 54–63. Association for Computational Linguistics.

Baumgartner, J., Zabbettiym Savvas, B. Keegan, M. Squire, and J. Blackburn (2020). The Pushshit Reddit Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media, 14(1)*, pp. 830–839.

Bhattacharya, S., S. Singh, R. Kumar, A. Bansal, A. Bhagat, Y. Dawer, B. Lahiri, and A. K. Ojha (2020, 3). Developing a Multilingual Annotated Corpus of Misogyny and Aggression. *arXiv preprint arXiv:2003.07428*.

Borkan, D., L. Dixon, J. Sorensen, N. Thain, and L. Vasserman (2019, 5). Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, New York, NY, USA, pp. 491–500. ACM.

Buciluǎ, C., R. Caruana, and A. Niculescu-Mizil (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, New York, New York, USA, pp. 535. ACM Press.

Burnap, P. and M. L. Williams (2015). Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. *Policy & Internet. Wiley Periodicals 7*(2), 223–242.

Cambridge University Press (2022). Definition of Hate Speech from the Cambridge Advanced Learner's Dictionary & Thesaurus. Available online at: https://dictionary.cambridge.org/dictionary/english/hate-speech. Accessed on: 04/03/2022.

Caselli, T., V. Basile, M. Jelena, K. Inga, G. Michael, et al. (2020). I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Language Resources and Evaluation Conference*, pp. 6193–6202. The European Language Resources Associatio.

Caselli, T., V. Basile, J. Mitrovic, and M. Granitzer (2021). HateBERT: Retraining BERT for Abusive Language Detection in English. *arXiv preprint arXiv:2010.12472*.

Chalkidis, I., M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos (2020). Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Chapman, P., J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth (2000). CRISP-DM 1.0: Step-by-step data mining guide.

Cohen, J. (1960, 4). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*(1), 37–46.

Davidson, T., D. Warmsley, M. Macy, and I. Weber (2017, 3). Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Volume 11.

De Gibert, O., N. Perez, A. García-Pablos, and M. Cuadros (2018). Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pp. 11–20.

DeepAI (2022). Accuracy (error rate). Available online at: https://deepai.org/machine-learning-glossary-and-terms/accuracy-error-rate. Accessed on: 07/04/2022.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018, 10). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Downs, A. (1972). Up and Down With Ecology: The "Issue-Attention Cycle". *The public 28*, 38–50.

Edwards, A. (2016, 6). UNHCR viewpoint: 'Refugee' or 'migrant' – Which is right? Available online at: https://www.unhcr.org/news/latest/2016/7/55df0e556/. Accessed on: 04/02/2022.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*(5), 378–382.

Fortuna, P., J. Soler-Company, and L. Wanner (2020). Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 6786–6794. European Language Resources Association (ELRA).

Founta, A.-M., C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Twelfth International AAAI Conference on Web and Social Media*, pp. 491–500.

Gao, L. and R. Huang (2018, 10). Detecting Online Hate Speech Using Context Aware Models. *In Proceedings of Recent Advances in Natural Language Processing*, 260–266.

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly Media, Inc.

Gokaslan, A. and V. Cohen (2019). OpenWebTextCorpus. Available online at: http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus. Accessed on: 12/05/2022.

Google (2020). Machine Learning Crash Course - Classification: Accuracy. Available online at: https://developers.google.com/machine-learning/crash-course/classification/accuracy. Accessed on: 07/04/2022.

Hasty visionAI Wiki (2021). Warm-Up. Available online at: https://wiki.hasty.ai/scheduler/warm-up. Accessed on: 02/04/2022.

Hinton, G., O. Vinyals, and J. Dean (2015, 3). Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* 2(7).

Hutto, C. J. and E. Gilbert (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the international AAAI conference on web and social media*, pp. 216–225.

IBM Cloud Education (2020, 5). Deep Learning. Available online at: https://www.ibm.com/cloud/learn/deep-learning. Accessed on: 01/05/2022.

Kennedy, B., M. Atari, A. M. Davani, L. Yeh, A. Omrani, Y. Kim, K. Coombs, S. Havaldar, G. Portillo-Wightman, E. Gonzalez, J. Hoover, A. Azatian, A. Hussain, A. Lara, G. Cardenas, A. Omary, C. Park, X. Wang, C. Wijaya, Y. Zhang, B. Meyerowitz, and M. Dehghani (2022, 3). Introducing the Gab Hate Corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation* *56*(1), 79–108.

Kirk, H. R., B. Vidgen, P. Röttger, T. Thrush, and S. A. Hale (2021, 8). Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate. *arXiv preprint arXiv:2108.05921*.

Koo, T., X. Carreras, and M. Collins (2008). Simple Semi-supervised Dependency Parsing. In *Proceedings of ACL-08: HLT*, pp. 595–603.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability. Retrieved from: https://repository.upenn.edu/asc_papers/43.

Kumar, R., A. K. Ojha, M. Zampieri, and S. Malmasi (2020). TRAC - 2. Available online at: https://sites.google.com/view/trac2/shared-task. Accessed on: 07/02/2022.

Kurrek, J., H. M. Saleem, and D. Ruths (2020). Towards a Comprehensive Taxonomy and Large-Scale Annotated Corpus for Online Slur Usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, Stroudsburg, PA, USA, pp. 138–149. Association for Computational Linguistics.

Kwok, I. and Y. Wang (2013). Locate the Hate: Detecting Tweets against Blacks. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Lai, G., Q. Xie, H. Liu, Y. Yang, and E. Hovy (2017, 4). RACE: Large-scale ReAding Comprehension Dataset From Examinations. *arXiv preprint arXiv:1704.04683*.

Law, H., B. Mugo, J. McKiernan, R. Vasquez, E. Feller, I. Khan, S. Jaquemet, P. Leclerc, A. B. Johnsson, C. Pintat, R. Huizenga, and K. Jabre (2001). *Refugee Protection: A Guide to International Refugee Law*. Office of the United Nations High Commissioner for Refugees [and] Inter-Parliamentary Union.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov (2019, 7). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. and F. Hutter (2017, 11). Decoupled weight decay regularization.

Lundberg, S. (2018). An introduction to explainable AI with Shapley values. Available online at: https://shap.readthedocs.io/en/latest/index.html. Accessed on: 30/04/2022.

Lundberg, S. M., P. G. Allen, and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems 30*, 4765–4774.

Madabushi, H. T., E. Kochkina, and M. Castelle (2020, 3). Cost-Sensitive BERT for Generalisable Sentence Classification with Imbalanced Data. *arXiv preprint arXiv:2003.11563*.

Mandl, T., S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, and A. Patel (2019, 12). Overview of the HASOC track at FIRE 2019. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, New York, NY, USA, pp. 14–17. ACM.

Mathew, B., R. Dutt, P. Goyal, and A. Mukherjee (2019). Spread of Hate Speech in Online Social Media. *In Proceedings of the 10th ACM conference on web science*, 173–182.

Mathew, B., P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee (2020, 12). HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289*.

Müller, K. and C. Schwarz (2021, 8). Fanning the Flames of Hate: Social Media and Hate Crime. *Journal of the European Economic Association 19*(4), 2131–2167.

Nagel, S. (2016). Cc-news. Available online at: http://web.archive.org/save/http://commoncrawl.org/2016/10/news-dataset-available. Accessed on: 12/05/2022.

Ng, A., Y. B. Mourri, and K. Katanforoosh (2021). Structuring Machine Learning Projects. Available online at: https://www.coursera.org/lecture/machine-learning-projects/transfer-learning-WNPap. Accessed on: 20/04/2022.

Nockleby, J. T. (2000). Hate Speech. Encyclopedia of the American Constitution.

Ousidhoum, N., Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung (2019). Multilingual and Multi-Aspect Hate Speech Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA, pp. 4674–4683. Association for Computational Linguistics.

PELTARION (2022). Macro F1-Score. Available online at: https://peltarion.com/knowledge-center/documentation/evaluation-view/classification-loss-metrics/macro-f1-score. Accessed on: 07/04/2022.

Pennington, J., R. Socher, and C. D. Manning (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Poletto, F., V. Basile, M. Sanguinetti, C. Bosco, and V. Patti (2021, 6). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation 55*(2), 477–523.

Rajbhandari, S., J. Rasley, O. Ruwase, and Y. He (2019, 10). ZeRO: Memory Optimizations Toward Training Trillion Parameter Models. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2020-November*.

Rajpurkar, P., R. Jia, and P. Liang (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Stroudsburg, PA, USA, pp. 784–789. Association for Computational Linguistics.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, pp. 2383–2392. Association for Computational Linguistics.

Ribeiro, M. T., T. Wu, C. Guestrin, and S. Singh (2020, 5). Beyond Accuracy: Behavioral Testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118*.

Ripley, B. D. (1996, 1). *Pattern Recognition and Neural Networks*. Cambridge University Press.

Roberts, C., M. Innes, M. Williams, J. Tregidga, D. Gadd, and N. Cook (2013). Understanding who commits hate crime and why they do it. *Welsh Government Social Research*. Available online at: https://orca.cardiff.ac.uk/id/eprint/58880/1/understanding-who-commits-hate-crime-and-why-they-do-it-en.pdf. Accessed on: 12/05/2022.

Röttger, P., B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. B. Pierrehumbert (2020, 12). HateCheck: Functional Tests for Hate Speech Detection Models. *arXiv preprint arXiv:2012.15606*, 41–58.

Samghabadi, N. S., P. Patwa, S. Pykl, P. Mukherjee, A. Das, and T. Solorio (2020). Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 126–131. European Language Resources Association (ELRA).

Sanguinetti, M., F. Poletto, C. Bosco, V. Patti, and M. Stranisci (2018). An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2799–2805.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2019, 10). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Saunders, M., P. Lewis, and A. Thornhill (2019). *Research Methods for Business Students* (8. ed.). Pearson Education Limited.

Schmidt, A. and M. Wiegand (2017). A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10. Association for Computational Linguistics.

scikit-learn (2022). Documentation: F1-Score. Available online at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. Accessed on: 07/04/2022.

Sekaran, U. and R. Bougie (2016). *Research Methods For Business: A Skill Building Approach* (7 ed.). John Wiley & Sons.

Stevis-Gridneff, M. and C. Gall (2020, 2). Erdogan Says, 'We Opened the Doors,' and Clashes Erupt as Migrants Head for Europe. Available online at: https://www.nytimes.com/2020/02/29/world/europe/turkey-migrants-eu.html. Accessed on: 04/05/2022.

Swamy, S. D., A. Jamatia, and B. Gambäck (2019). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pp. 940–950.

Taylor, W. L. (1953, 9). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly 30*(4), 415–433.

Trinh, T. H. and Q. V. Le (2018, 6). A Simple Method for Commonsense Reasoning. *arXiv preprint arXiv:1806.02847*.

UNHCR (1998, 7). Guiding Principles on Internal Displacement. Technical report, New York. Available online at: https://www.refworld.org/docid/3c3da07f7.html. Accessed on: 04/02/2022.

UNHCR (2006, 6). Global Report 2005. Technical report, Geneva. Available online at: https://www.unhcr.org/publications/fundraising/4a0c04f96/global-report-2005.html. Accessed on: 04/02/2022.

UNHCR (2010, 12). *Convention and Protocol related to the Status of Refugees*. Geneva. Available online at:https://unhcr.org/3b66c2aa10. Accessed on: 11/04/2022.

UNHCR (2011, 4). Protection Training Manual for European Border and Entry. Technical report, Brussels. Available online at:https://www.unhcr.org/4d944c319.html. Accessed on: 26/01/2022.

UNHCR (2016, 6). Guidelines on International Protection No. 12. Technical report. Available online at: www.refworld.org/docid/583595ff4.html. Accessed on: 12/05/2022.

UNHCR (2019). COUNTERING TOXIC NARRATIVES ABOUT REFUGEES AND MIGRANTS. Available online at:https://www.unhcr.org/5df9f0417.pdf. Accessed on: 24/01/2022.

UNHCR (2021). MID-YEAR TRENDS 2021. Technical report. Available online at: https://www.unhcr.org/statistics/unhcrstats/618ae4694/mid-year-trends-2021.html. Accessed on: 12/05/2022.

UNHCR (2022a, 4). Afghanistan Situation Update. Technical report. Available online at: https://data2.unhcr.org/en/documents/details/92260. Accessed on: 04/05/2022.

UNHCR (2022b). Situation Afghanistan. Available online at: https://data2.unhcr.org/en/situations/afghanistan. Accessed on: 04/05/2022.

United Nations (2019, 6). UN Strategy and Plan of Action on Hate Speech - Overview. Technical report. Available online at: https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf. Accessed on: 09/11/2021.

United Nations (2020, 9). United Nations Strategy and Plan of Action on Hate Speech. Technical report. Available online at: https://www.un.org/en/genocideprevention/hate-speech-strategy.shtml. Accessed on: 09/05/2022.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017, 12). Attention Is All You Need. *Advances in neural information processing systems*.

Vidgen, B. and L. Derczynski (2020, 12). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one 15*(12), e0243300.

Vidgen, B., D. Nguyen, H. Margetts, P. Rossini, and R. Tromble (2021). Introducing CAD: the Contextual Abuse Dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, USA, pp. 2289–2303. Association for Computational Linguistics.

Vidgen, B., T. Thrush, Z. Waseem, and D. Kiela (2020, 12). Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection. *arXiv preprint arXiv:2012.15761*, 1667–1682.

Vig, J. (2019). A Multiscale Visualization of Attention in the Transformer Model. pp. 37–42.

Wang, A., A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman (2019, 4). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

Warner, W. and J. Hirschberg (2012). Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pp. 19–26. Association for Computational Linguistics.

Waseem, Z. and D. Hovy (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of NAACL-HLT*, pp. 88–93.

Wiegand, M., J. Ruppenhofer, and T. Kleinbauer (2019). Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North*, Stroudsburg, PA, USA, pp. 602–608. Association for Computational Linguistics.

Williams, M. L., P. Burnap, A. Javed, H. Liu, and S. Ozalp (2020, 1). Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology 60*(1), 242–242.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush (2019, 10). HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Wolters, J.-N. and N. Olšavský (2021, 9). Framing the Refugee Debate. Investigation of Frames on Twitter in Response to Refugee-related Events. Technical report, Copenhagen Business School.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean (2016, 9). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Wulczyn, E., N. Thain, and L. Dixon (2016). Wikimedia Detox. Available online at: https://meta.wikimedia.org/wiki/Research:Detox/Data_Release. Accessed on: 07/03/2022.

Wulczyn, E., N. Thain, and L. Dixon (2017, 4). Ex Machina. In *Proceedings of the 26th International Conference on World Wide Web*, Republic and Canton of Geneva, Switzerland, pp. 1391–1399. International World Wide Web Conferences Steering Committee.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French. *arXiv preprint cmp-lg/9406034*, 88–95.

Zampieri, M., S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar (2019, 2). Predicting the Type and Target of Offensive Posts in Social Media. *arXiv preprint arXiv:1902.09666*.

Zhu, J., Z. Tian, and S. K. Ubler (2019). UM-IU@LING at SemEval-2019 Task 6: Identifying Offensive Tweets Using BERT and SVMs. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*, pp. 788–795. Association for Computational Linguistics.

Zhu, Y., R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler (2015, 12). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27. IEEE.

A APPENDIX

A.1 FIGURES

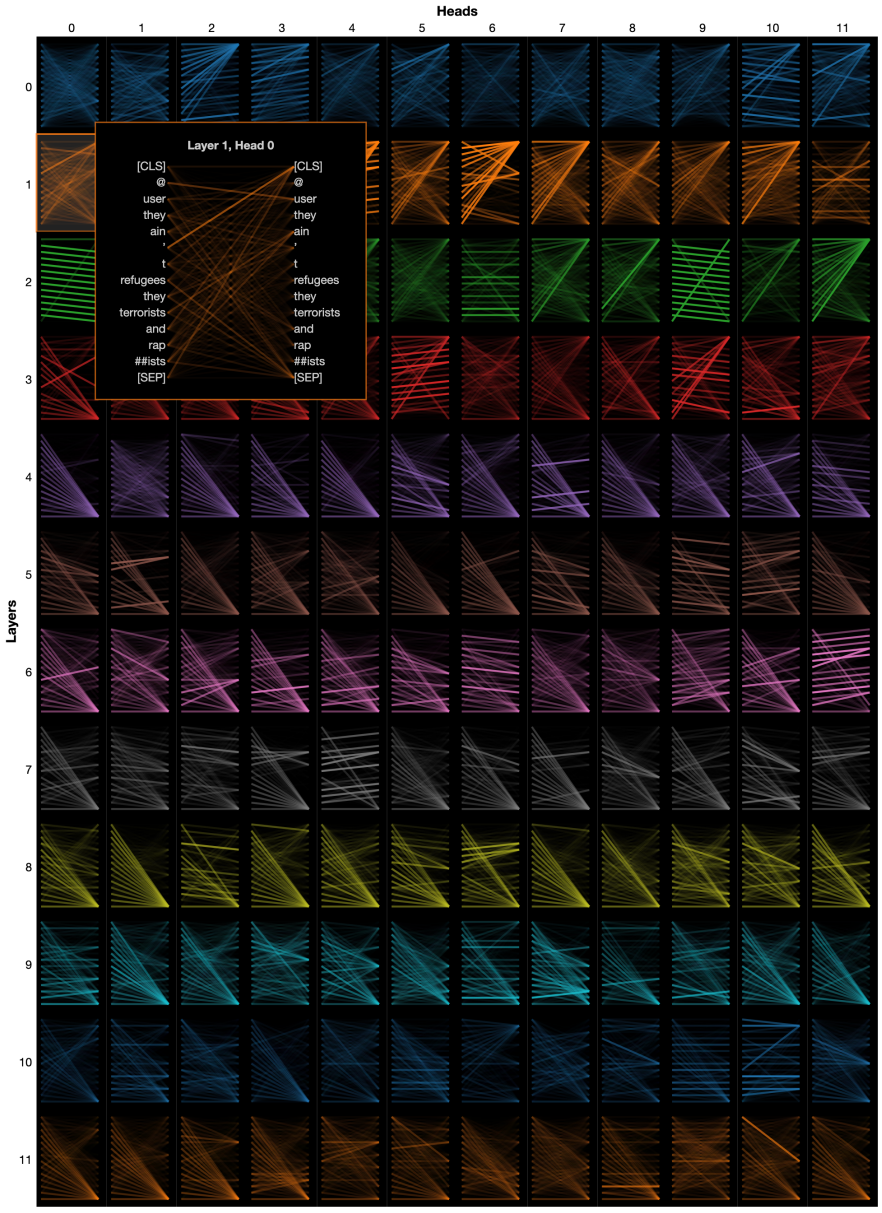

Figure 21: Model view representation of the HateBERT model for each layers' attention head for the tweet *"@USER they ain't refugees they terrorists and rapists"* based on Vig (2019).

A.2 TABLES

dataset	count	mean	std	min	25%	50%	75%	max
cad	1037	41	84	1	9	20	42	1779
civil	195642	47	43	1	16	33	64	311
davidson	24783	14	7	1	9	13	19	52
dynhs	41144	24	25	1	9	16	31	402
ghc	27546	21	17	1	9	15	28	268
hasoc	7005	24	13	2	14	22	34	95
hatemoji	5912	8	5	1	5	7	10	51
hateval	10000	22	11	1	13	20	27	93
hatexplain	20148	24	14	2	12	21	34	165
ousid	5647	9	4	1	5	9	12	24
slur	40000	34	25	1	16	27	44	167
wikipedia	93755	48	48	1	14	31	63	256

Table 35: Statistical overview of sentence word lengths for each dataset.

Token	Example subset of text originally containing token
igger	"*nigger", "***trigger", "before;nigger", "currynigger", "sandnigger"
igger·	"nigger-lover′"", "trigger?", "trigger′", "niggers".", "diggers"
black	"blacklock,", "blackness", "earlier.....blacks", "blacker"
agg	"faggot...", "aggies", "teagger", "ragged", "self-aggrandizing"
ike	"strikes:", "like.′"", ""likes",", "strikes,activists", "child-like"
lam·	"""islam" "selam", "lam", "banislam", ""islam", "slam"
gay·	"antigay", "turgay", "mgay", "antiiiiiii-gay", "supergay", "pro-gay"
anny	"tyranny...why", "shannyn,", "tranny.)", "granny?", "tranny."", "fag/tranny"
oes	"doesen′t", "negroes,", "embargoes", "egoes"
ches·	""bitches", "watches", "leaches", "turkroaches", "churches"
ches	"chess,", "bitches.", "turkroaches", "sandwiches", "trenches"
hypocr	"hypocrites??", "hypocritical′", "hypocrites.....and", "hypocrisy.."
refugees·	"4,000+refugees", "ussendbackafghanrefugees", "openeuborderstorefugees"
refugees	"refugees.millions", "refugeeswelcome", "refugees.@bjp4india", "″refugees″...."
migrants	"migrants.", "immigrants!!", "immigrants", ""migrants", "immigrants","
ref	"displaced/refugee", "turkey.refugees", "dehumanised.refugeesgr", "refugee.....u"
rap	"arapaio", "rapeculture", "bootstrap", "goat-rapists"
rians	"missourians", "afghans,africans,algerians", "we,tigrians", "identitarians"
ylum	"fraud>asylum", "welcome„asylum", "seekersbrexitasylum"
lim	"climatic", "sublime", "limited...like", "limitless.", "muslim-american."
uge	"alone-huge", "refugees,:Pakistan:", "refugees. . . there", "less*refugees"
ians	"theologians,", "syrians.′", "mountians"
ians·	"nigerians", "christians", "skying,ethiopians", "wantsyrians", "“politicians"
rica	"americans...", "africa... 😂 😂 ", "america!?"
uge	"refugees.....turkey", "refugees.kick", "to′refugees′got", "refugees&the"
ors	"′worst-ever", "censorship?", "actors"", "radiators.", "neighbors:"
bers	"outnumbers", "bombers"", "chambers,and"
anal	"organal", "analogical,", "analysis′"", "analyticism"
eker	"seekers.truthfully", "seekers...″"", "seekers,"", "seeker...", "seeker/refugee."
iest	"priests′", "healthiest", "wealthiest.", "countriesturkeyiranarab", "charliestayt"
itto	"unfittobepresident", "(ditto)", "vittorio",

Table 36: Tokens with selected examples of their usage from the combined test set, UNHCR refugee data, and all refugee crises.

A.3 REFUGEE CRISES

A.3.1 United Kingdom Channel Crossing Refugee Crisis

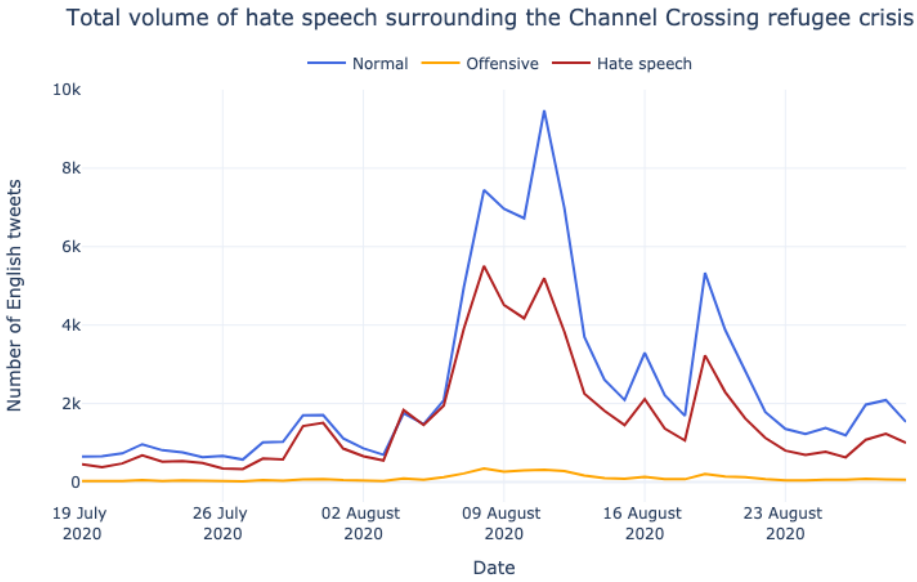

Figure 22: Development of the volume of hate, offensive, and normal speech in the period from 19-07-2020 to 29-08-2020 for the Channel Crossing refugee crisis predicted with RoBERTa.

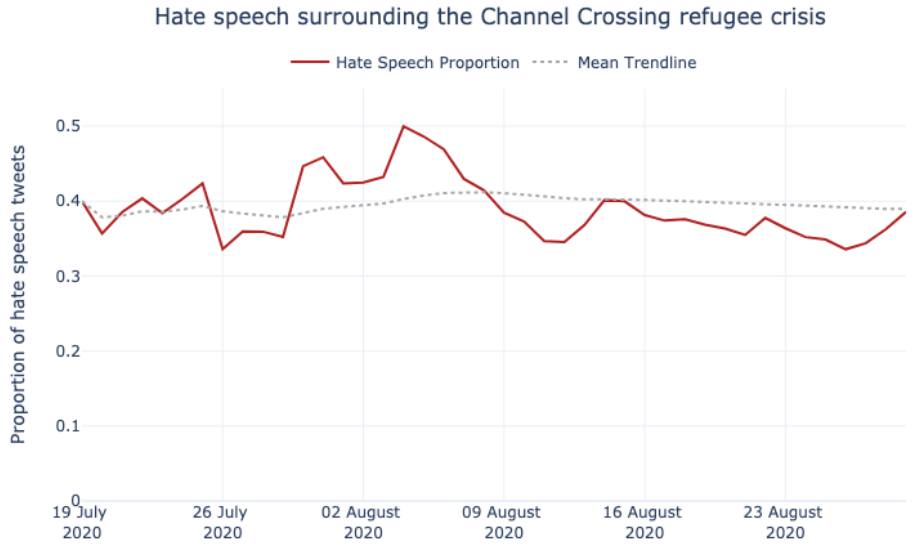

Figure 23: Development of the proportion of hate speech in % in the period from 19-07-2020 to 29-08-2020 for the Channel Crossing refugee crisis.

Token	n	Hate speech
immigrants	13	0.30
migrants	16	0.24
immigrants·	31	0.23
migrants·	60	0.20
migrant	2	0.17
igrants·	3	0.17
refugees·	33	0.16
towards·	2	0.16
ylum·	2	0.15
seekers	4	0.15
refugees	16	0.14
migrant·	18	0.14

Table 37: Feature importance of tokens with more than 1 occurrences sorted by hate speech for the Channel refugee crisis. SHAP values are calculated using the mean of a subset of the dataset. Whitespace after a token is marked with a middle dot (·) indicating that the token is followed by another word.

A.3.2 Rohingya Refugee Crisis

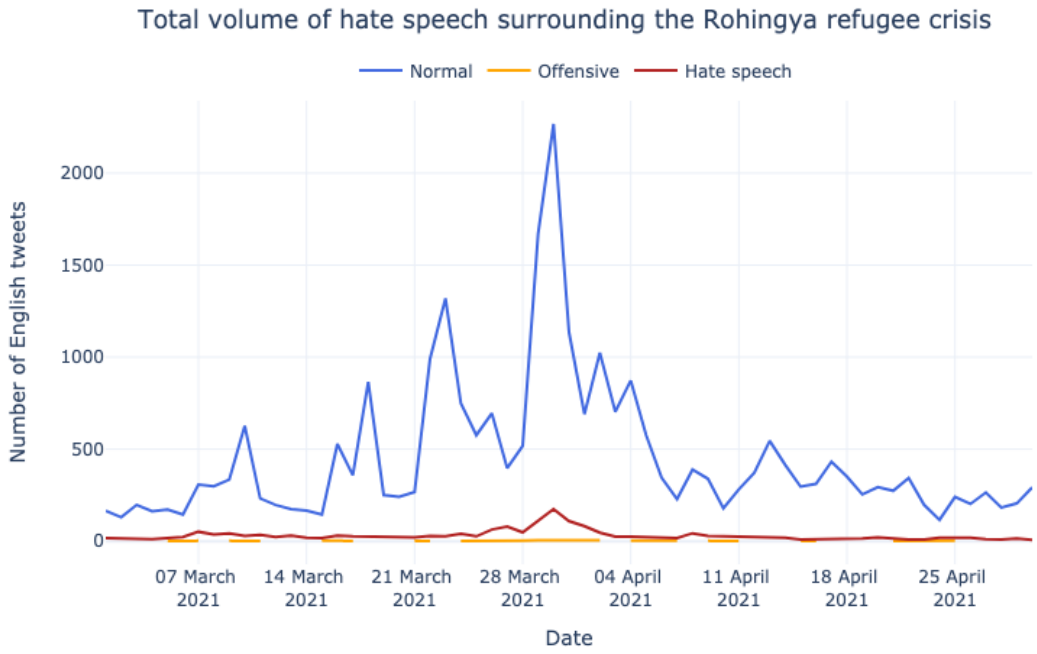

Figure 24: Development of the volume of hate, offensive, and normal speech in the period from 01-03-2021 to 30-04-2021 for the Rohingya refugee crisis predicted with RoBERTa.

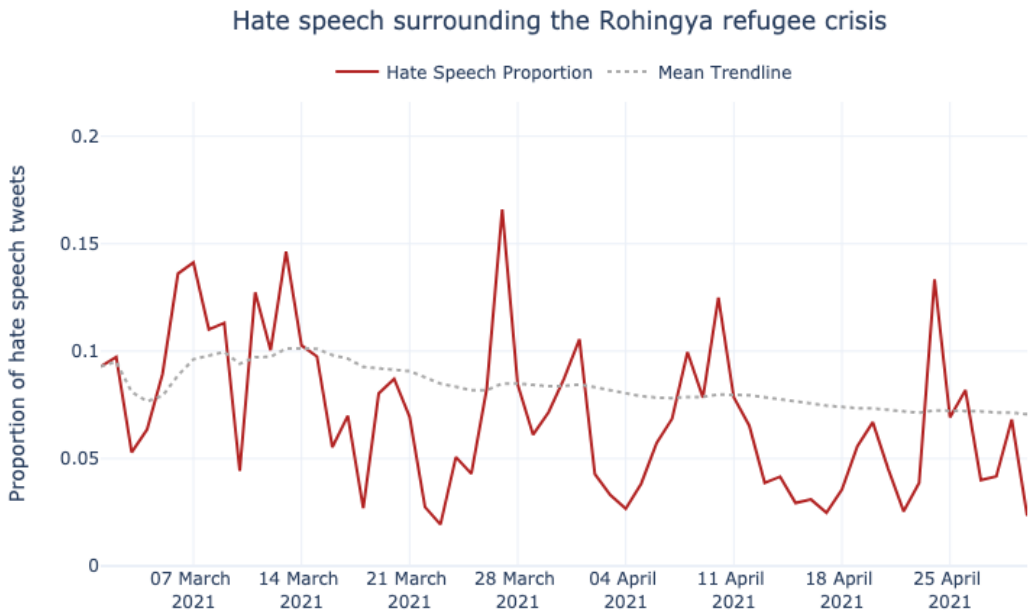

Figure 25: Development of the proportion of hate speech in % in the period from 01-03-2021 to 30-04-2021 for the Rohingya refugee crisis.

Token	n	Hate speech
refugees	13	0.12
illegal·	4	0.10
workers·	2	0.08
immigrants·	2	0.08
migrant·	2	0.08
refugees·	43	0.07
ians	2	0.06
anmar	10	0.06
uge	17	0.05
refugee·	28	0.05
lim	3	0.05
igrants·	2	0.05

Table 38: Feature importance of tokens with more than 1 occurrences sorted by hate speech for the Rohingya refugee crisis. SHAP values are calculated using the mean of a subset of the dataset. Whitespace after a token is marked with a middle dot (·) indicating that the token is followed by another word.

A.3.3 Tigray Refugee Crisis

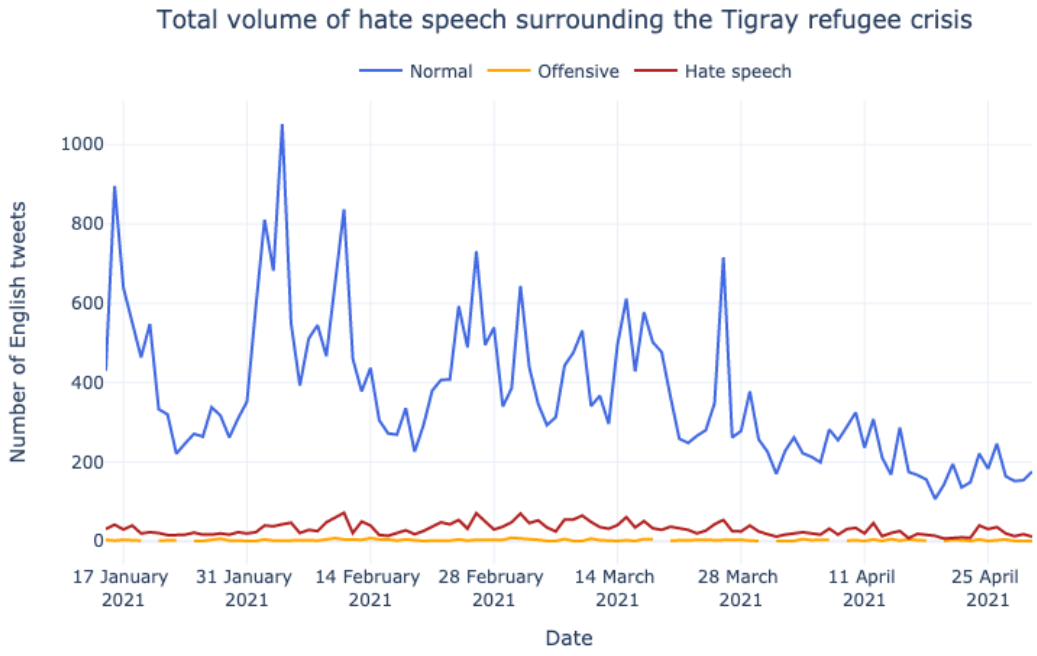

Figure 26: Development of the volume of hate, offensive, and normal speech in the period from 15-01-2021 to 30-04-2021 for the Tigray refugee crisis predicted with RoBERTa.

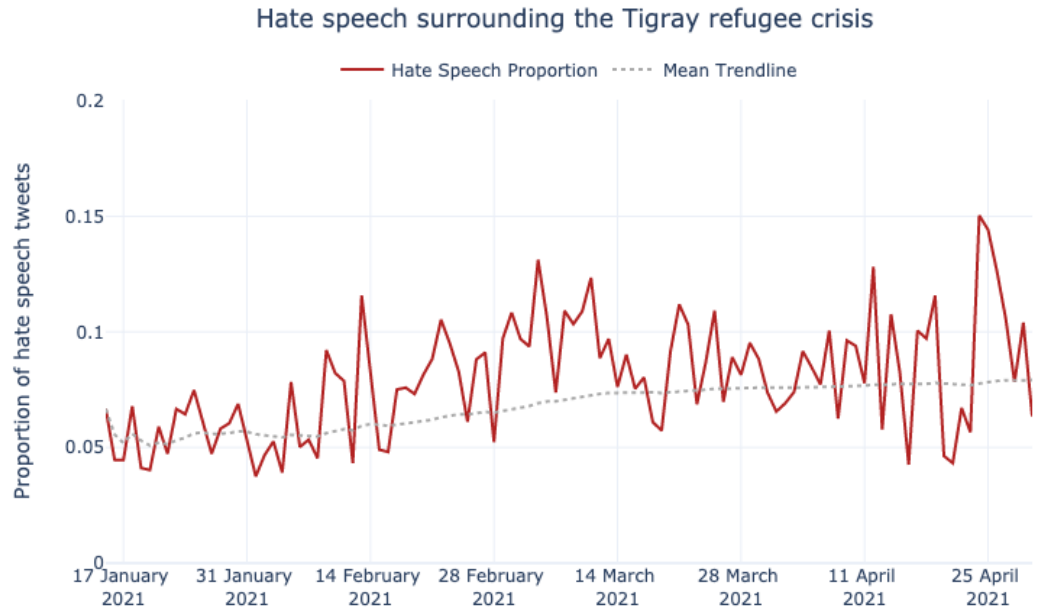

Figure 27: Development of the proportion of hate speech in % in the period from 15-01-2021 to 30-04-2021 for the Tigray refugee crisis.

Token	n	Hate speech
refugee	3	0.21
immigrants	2	0.15
immigrants·	2	0.14
migrants	4	0.13
refugees	8	0.11
migrants·	2	0.11
rica	2	0.08
ians·	3	0.07
refugees·	60	0.07
migrant·	3	0.07
uge	45	0.07
refugee·	19	0.06

Table 39: Feature importance of tokens with more than 1 occurrences sorted by hate speech for the Tigray refugee crisis. SHAP values are calculated using the mean of a subset of the dataset. Whitespace after a token is marked with a middle dot (·) indicating that the token is followed by another word.