MASTER THESIS

Multilingual Racial Hate Speech Detection: Annotation and Experimentation

vorgelegt von

Skadi Jule Dinter

MIN-Fakultät

Fachbereich Informatik

Studiengang: Informatik

Matrikelnummer: 6948034

Erstgutachter: Prof. Dr. Chris Biemann

Zweitgutachter: Dr. Seid Muhie Yimam

Zusammenfassung

Hatespeech, *Hassrede* oder *Hetze*, ist durch die Verbreitung von sozialen Medien und der einfachen Äußerung von Gedanken zu einem weit verbreiteten Phänomen geworden. Dadurch können sowohl auf individueller als auch auf gesellschaftlicher Ebene verheerende Konsequenzen entstehen. Häufig gibt es einen Zusammenhang zwischen Kriminalität oder Terror sowie Hassäußerungen im Netz.

In dieser Arbeit werden rassistische Tweets aus dem Monat nach dem Tod von George Floyd im Mai 2020 auf Französisch, Deutsch und Amharisch analysiert. Dieses Ereignis und die folgenden Diskussionen haben zu einem Anstieg in rassistischen Tweets in vielen Ländern geführt. Mittels Stichwortlisten wurden Tweets vorgefiltert und anschließend mittels Crowdsourcing in eine der folgenden Kategorien kategorisiert: Hass, Beleidigung, normale Sprache oder unsicher. Für die ersten beiden Kategorien sollte zudem noch angegeben werden, ob es ein rassistisches Ziel gibt. Die Datensätze sowie die Personengruppe, die die Tweets klassifiziert haben, wurden analysiert. Der französische Datensatz hat einen ausreichenden Inter-annotator-agreement-score (Fleiss Kappa) und konnte daher für die Erstellung von Hatespeech-Erkennungsmodellen genutzt werden. Diese basieren auf vortrainierten BERT Modellen sowie dem übersetzten Datensatz von HateXplain und wurden spezifisch für diese Aufgabe trainiert. Die Erkennung von normaler Sprache und Beleidigung funktioniert bereits für die meisten Tweets, aber es bestehen noch falsche Klassifizierungen (besonders, um die Kategorie Hass richtig zu erkennen).

Abstract

The social media have made it possible to express sentiments freely, including racist utterings or hate speech which are now widespread. As hate speech can have severe consequences both on an individual level and on the society, it is important to combat this problem. For example, there are often connections between hate crimes and hatespeech that was posted online.

In this work, racist tweets that were published after the death of George Floyd in May 2020 are analysed. This event has accelerated protests, debates around racism as well as racist utterings in many countries. Tweets that were published in the month after the death are analysed in three different languages: German, French and Amharic. The tweets are filtered by keyword lists and then annotated with crowdsourcing tools as hate, offensive, normal speech or unsure. In case one of the first options is chosen, the annotators choose if the tweet is directed against a racial target or not or if they are unsure. The demographics of the annotators as well as the annotations are analysed.
The French dataset has a sufficient inter-annotator-agreement score (Fleiss Kappe) and is therefore used to build hate speech detection models. These are based on BERT models and the translated HateXplain dataset. Different experiments show that the models are able to predict normal and offensive speech for most of the tweets. There remain several misclassifications, especially to detect hate speech. Thus, a hate speech detection model based on BERT and HateXplain could be efficiently adapted to French.

Disclaimer

Due to the content of this thesis, it contains possibly vulgar, hurtful or offensive language. Quoting this is useful for explanatory purposes but does not reflect in any way the opinions of the author.

Acknowledgements

I would like to acknowledge the support of various people which helped to make this thesis possible. I am particularly grateful for the assistance given by Dr. Seid Muhie Yimam who supported me in many ways, was always available and gave valuable feedback and consultations throughout the whole process of the project. Both him and Prof. Dr. Chris Biemann, who helped me with lots of great suggestions on how to tackle the thesis, were supporting me in numerous ways and made this thesis possible. Abinew Ali Ayele helped a lot with the Amharic language: by preprocessing and annotating the datasets as well as with the cultural context. Punyajoy Saha helped with the architecture setup for the models. Last but not least I want to thank all persons proof-reading the thesis for taking the time and giving very helpful feedback.

Contents

1. Introduction

In this chapter I will introduce hate speech in social media as a global and challenging problem. The increasing amount of hate speech has severe consequences both on an individual and societal level which motivates different attempts to combat hate speech. The automatic hate speech detection and their challenges are outlined. The research question is introduced.

1.1. Motivation

The rise of social media platforms like Facebook or Twitter in the last years has enabled users to express and distribute their sentiments on events, ideas, products or other persons freely and conveniently. This eases the usage of hateful messages which can imply threats or harassment (Chiril, Benamara et al. 2019). Hateful content in various kinds has already been a long-existing problem but the immediacy and usually larger audience in social media have simplified the spreading of hate messages tremendously (Seoane and Monnier 2019). **Hate speech** as a specific form of hateful content is not universally defined and agreed upon as it depends on the personal intention and perception (Schröder 2020).

Hate speech is a public communication with direct attacks or violence based on a specific characteristic, this will be discussed in Chapter 2. Hate speech is usually motivated by prejudices or disgust of a group of persons. There are a lot of different ways to express hate speech.
Offensive speech can be very similar to hate speech as it is hurtful speech directed against another person but expressing this kind of speech has fewer legal implications.
Racism as a type of discrimination makes up a large portion of hate speech and is usually directed against the perceived ethnicity, appearance, religion or culture as described in Section 2.2. After the killing of George Floyd on May 25th, 2020 the amount of racist comments on social media platforms increased even more (see Section 2.2.1). Hate speech is a widespread problem in general; around 14.3% of all messages are perceived aggressive or hateful[1]. Racist and sexist content are especially widespread (Waseem 2016). These hateful posts can inspire hate crimes in the physical world globally[2]. It can also distort the public perception of opinions and thus divide a society. Those affected from the posts as well as witnesses tend to isolate themselves from the public or political discussions leading to an even larger separation of the society. Additionally, the victims tend to suffer from problems both psychologically and at their work or education[3].

In Chapter 2, the specific situation regarding hate speech, the legal framework and connection to the physical world are described for the following examplary countries in which the analysed

[1] Cartographie de la haine: https://www.isdglobal.org/wp-content/uploads/2020/01/Cartographie-de-la-haine-fr.pdf
[2] Hate Speech on Social Media: Global Comparisons, Zachary Laub: https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons#chapter-title-0-2
[3] Institut für Demokratie und Zivilgesellschaft, Geschke et. al.: https://www.idz-jena.de/forschung/hass-im-netz-eine-bundesweite-repraesentative-untersuchung-2019/

languages are spoken: Germany, France, the United States and Ethiopia. Furthermore, the consequences described above are detailed.

In the context of these severe consequences, the social media platforms and academia have tried to improve the detection of hate speech in order to advance the treatment of reported posts (Waseem 2016). In academia, there is a strong interest in detecting types of hate and there are many workshops and challenges designated to this topic (Mozafari, Farahbakhsh and Crespi 2020). Currently, the social media platforms use mainly content moderation systems which are human-machine collaborative systems[4], to detect and handle hate speech. This way, both humans and machines check content to decide if it aligns with the platform's guidelines and the laws of the corresponding country in which the content was published.

Classification challenges When detecting hate speech automatically with a classification algorithm, a lot of difficulties arise due to the nature of hate speech. The posts will not even always be classified as the same category by humans due to their different perspectives (Alkomah and Ma 2022). The hate may be hidden in irony and the context or semantic can change rapidly requiring continuous learning of the algorithm. The hate speech classification research has been done mostly for English language tweets, although only 32% of the posts are written in English[5], followed by Japanese with 19%. Around 3% of the tweets are written in French[5].

Especially low-ressource languages like Amharic (mainly spoken in Ethiopia) have not been researched extensively so far. This is due to limited resources or conflicting interests (Pohjonen and Udupa 2017). As the perception of hate speech can be different depending on the language or culture, it is necessary to further develop the automatic detection in low-ressource languages to understand the corresponding contexts and to improve the classification (Salminen, Veronesi et al. 2018).

1.2. Research question

As described in Section 1.1, hate speech detection is usually limited to the English language. In order to advance the development of hate speech detection algorithms in multiple languages, I want to extend the English hate speech detection model HateXplain (Mathew et al. 2021) to German, French and Amharic. As racism is a very common issue in today's society (see Section 2.2) and due to the limited scope of this thesis, I place the emphasis on detecting racial hate speech. Other forms of discrimination as well as intersectionality will need to be considered in future research.

For the experimental set-up, public tweets are collected from the 1% Twitter stream[6], randomly selecting 1% of all tweets published. Their language is detected with the python package pycld2[7]. Tweets written only in German, in French or Amharic are further preprocessed and annotated

[4] Online content moderation, Dr. Savvas Zannettou: https://www.mpi-inf.mpg.de/departments/inet/online-content-moderation

[5] 2018 Research on 100 Million Tweets, Vicinitas:https://www.vicinitas.io/blog/twitter-social-media-strategy-2018-research-100-million-tweets#language

[6] Twitter 1% sampled stream https://developer.twitter.com/en/docs/twitter-api/tweets/volume-streams/introduction

[7] pycld2: https://pypi.org/project/pycld2/

by crowdworkers. These annotators classify the tweet into hate speech or offensive speech and indicate if it contains racial speech. Possible annotation biases are considered.

These annotations will be used to build corresponding hate speech detection models. The baseline dataset is HateXplain (Mathew et al. 2021). It is extended to French using the collected and annotated dataset for training. The model is trained to detect racial hate speech.

The classifications from the model will be analysed and language and culture-specific features are considered. Additionally, possible reasons for linguistical or cultural differences are explored.

By this work the following question is intended to be answered:

> **Can we create a racial hate speech detection model (based on BERT and HateXplain) that can be efficiently adapted to other languages or cultures, specifically German, French and Amharic?**

By conducting this thesis the research on hate speech detection should be advanced. Several specificities should be pointed out that need to be considered when developing algorithms for further languages as well as other cultural or linguistical regions.

Structure of the thesis

In this chapter I have introduced the challenging global problem hate speech together with possible consequences and different ways to combat it like content moderation systems. In the Chapter 2 I will further explain hate speech and the situation in the countries I analyse. Related works will be presented in Chapter 3.1. Following this, the data collection and annotation process will be explained in detail in Chapter 4.1. Afterwards the model together with experiments are presented (see Chapter 5.1). In the discussion (Chapter 6.1) my approach and findings are presented and an outlook is given on future research in Chapter 7.

2. Background

This chapter will be used to develop a definition of hate speech and offensive speech together with its context in social media to better understand the challenges in the detection. Then, several examples of different kinds of hate speech are introduced. The consequences of hate speech are explained to motivate the research on hate speech detection. I will further explain the focus on racism and the situation in the different countries Germany, France, UK and Ethiopia. Several approaches to detect hate speech as well as their limitations are further outlined.

2.1. Hate speech

Hate speech as a specific kind of negative speech is not universally defined and agreed upon as it highly depends on the personal intention and perception as well as the context (Schröder 2020). As an example, terms like "n*gga *sic*" are commonly used to express hate.
However, people belonging to African American communities use these terms in a non-offensive way (Mozafari, Farahbakhsh and Crespi 2020).

In common parlance, hate speech is described as any kind of discriminating or attacking speech against certain people or groups, usually expressed in social media (Schröder 2020).

The following definition (Nockleby 2000) is often used in academia:

> Hate speech is commonly defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic.

The Cambridge Dictionary[1] specifies the communication to be public:

> public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation

The social media platforms define their regulations based upon the UN Declaration of Human Rights of 1948 (Siapera 2019).

Twitter is one of the largest social media platforms encountering hate speech, as there are around 230 million active daily users on Twitter[2]. It is studied extensively in this thesis and in their hateful conduct there are more characteristics named[3]:

[1] Definition of hate speech `https://dictionary.cambridge.org/dictionary/english/hate-speech`
[2] Number of monetizable daily active Twitter users (mDAU) worldwide from 1st quarter 2017 to 1st quarter 2022, Statista: `https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/`
[3] Twitter hateful conduct policy: `https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy`

> You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.

For the scope of this thesis, hate speech is defined as: **Public communication which may include direct attacks, violence or threats towards a person or a group based on a characteristic like: race, color, ethnicity, national origin, nationality, gender (identity), sexual orientation, sex, religion, caste, age, disability, or serious disease.**

Hate speech can be expressed in numerous ways including rhetoric figures making it more difficult to be identified. The examples in Table 2.1 illustrate possible forms of hate speech, in this case mainly in the form of sexism and racism (Chiril, Benamara et al. 2019).

Irony for example makes it difficult to detect the true intent of a speaker as they require attention from the reader to not being misinterpreted. This kind of hiding hate is very widespread, for example in 11% of hateful Italien tweets studied by Vidgen and Derczynski (2020). It appears often in self-deleted tweets and is commonly a cause for re-moderation. The annotation of ironic or sarcastic hateful content is very difficul and results in low inter-annotator agreements.

2.1.1. Problematic speech

Online harassment is the intention to intimidate, annoy or frighten a person online[4] by pursuing an unwanted and negative contact with a person. The perpetrator can be known by the victim or a stranger. The harassment does not necessarily need to be public[19].

Despite the definitions, only 37% in the US of those affected actually consider their experience to be harassment[19]. Thus, it is very difficult to handle these forms of hate when they are perceived very subjectively, especially from those affected.

Offensive speech

Offensive speech occurs very often in social media, too, and can be defined as „hurtful, derogatory or obscene comments made by one person to another person “(Bai et al. 2018). As hate speech has more severe legal and moral implications in contrast to offensive speech, it is important to distinguish the two even if it this might be difficult (Davidson, Warmsley et al. 2017).

Hate speech and offensive speech can be expressed in various ways. They also occur very often: according to Gatewood et al.[5], on a global scale 14.3% of the messages posted on social media platforms are perceived aggressive or hateful. More than half (53%) of the insults or aggressive comments are directed against other users, 30,1% against politicians, 15,5% against celebrities and 15.1% against media or journalists. 14% of people with immigration families have been affected by hate speech online[6].

[4] Online Harassment 2017, Pew Research Center: https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

[5] Cartographie de la haine: https://www.isdglobal.org/wp-content/uploads/2020/01/Cartographie-de-la-haine-fr.pdf

[6] Institut für Demokratie und Zivilgesellschaft, Geschke et. al.: https://www.idz-jena.de/forschung/hass-im-netz-eine-bundesweite-repraesentative-untersuchung-2019/

Form	Example	Translation	Reference
Negative opinion, abusive message	Meuf tu connais rien au foot. Tais toi. Contente de fan girler sur les joueurs et de mouiller sur MBappé	Girl, you do not know nothing about soccer. Shut up. Just fan girling about the players and wetting yourself on MBappé.	Chiril, Benamara et al. (2019)
Racial stereotype	Illegals are dumping their kids heres o they can get welfare, aid and U.S School Ripping off U.S Taxpayers #SendThemBack ! Stop Alowing illegals to Abuse the Taxpayer #Immigration.	-	Chiril, Benamara et al. (2019)
Humor, irony (often with naivety)	Le fait maison c'est toujours mieux. La preuve, on préfère toujours sa femme à sa prostituée. #humour.	Home-made is always better. To prove, we always prefer our wife to our prostitute. #humor.	Chiril, Benamara et al. (2019), Schröder (2020)
Benevolent sexism	Elle court vite pour une femme.	she runs fast for a women.	Chiril, Benamara et al. (2019)
Direct offense	miese GEZ Hure	lousy GEZ whore	Schröder (2020)
Announcement or support of violence	Bereite [...] dich [...] auf deine Hinrichtung vor	Prepare yourself for your execution	Schröder (2020)
Spread of fake news, myths	Die Flüchtlinge haben alle teure Handys	The refugees all have expensive smart phones	Schröder (2020),Maria Constantinou (2021)
Implicitly using sarcasm	Ich will auch ein neues Smartphone. Werd' ich im nächsten Leben halt Asylant.	I also wanna have a new smart phone. I am gonna be an asylum seeker in my next life then.	Maria Constantinou (2021), Schröder (2020)
Differentiating between them/us	unsere Frauen müssen vor denen geschützt werden	our women must be protected from those	Schröder (2020)
References (euphemisms or codes with a hateful meaning)	"Skype", "Google", and "banana"	-	Seoane and Monnier (2019), Yin and Zubiaga (2021)

Table 2.1.: Different kinds of hatespeech with examples

2.1.2. Consequences of Hate Speech

Hate messages are usually anonymously spread without facing the recipient and perceiving the consequences personally. They are often connected to violence in the physical world and can be a threat to democratic values (Schröder 2020).

On an individual level, exposure to hate speech online can have severe psychological effects (Schröder 2020). Usually those who are affected isolate themselves due to threats and stop or reduce their participation in online discourses. They can develop mistrust, anxieties or depressions with severe consequences. Among people younger than 25 almost the half of those affected by hate speech encounter emotional stress, more than every third encounters anxiety or depressions. For 15% of those who responded, problems at work or education arise. Female participants report these consequences more often[6]. Another way to deal with hate directed against oneself is to develop aggressive behaviour towards others which can also lead to or enforce violence and harm others (Schröder 2020). Additionally, racially offended people tend to experience immediate physical damage like head ache (Schröder 2020). Feelings like injustice or helplessness are often experienced[7].

Hate speech also affects witnesses (which are 40% of those interviewed, 73% of those 18-24 old): approximately only half of the users demonstrate their political opinion or participate in online discussions when encountering hate speech[6]. The consequence is that not only affected people but also those who already witnessed hate get pushed away – or they distance themselves – from discussions. This is to avoid to be affected themselves[6].

Furthermore, this type of hate has been the starting point for severe violence acts in the physical world such as terrorism, war or prohibition of free speech (Schröder 2020), see Section 2.3.

2.2. Racism

The rejection due to the perceived ancestry or skin color is a very common type of discrimination, particularly in the US (Boutwell et al. 2017). Among teenagers in the US, 52% witness often or sometimes racial hate speech[8]. It is very difficult to agree upon a universal definition of racism and racist utterances as these can be communicated, perceived and judged very differently depending on the person themself (Tulkens et al. 2016). There is also a lot of data missing as victims often do not report racist utterances because in their experience reporting does not solve the problem (Tulkens et al. 2016). In the UK, 76% of the hate crimes were racially motivated and in general, the hate crimes have more than doubled from 2013 to 2019 (Kilvington 2021). In the US, 61,8% of the hate crimes[9] (with only one targeted group) were racially motivated in 2020.

In order to include racial hate speech in this thesis, the definition from (Tulkens et al. 2016) is used considering social and cultural aspects additionally to physical or ethnic attributes:

[7] The dynamics of hate speech and counter speech in the social media, Centre for Internet and Human Rights, Katazyna Bojarska

[8] Encountering racial hate speech in the US https://www.statista.com/statistics/945392/teenagers-who-encounter-hate-speech-online-social-media-usa/

[9] FBI Releases 2020 Hate Crime Statistics https://www.justice.gov/hatecrimes/hate-crime-statistics#piechart

all negative utterances, negative generalizations and insults concerning ethnicity, nationality, religion and culture.

2.2.1. Racial Hate Speech example: George Floyd

On the May 25th, 2020, a video showing the killing of George Floyd by a police officer in Minneapolis, US, went viral. Following this happening, there were many racist postings but also protests against racism in general and policing with a racial bias (Priniski et al. 2021). The protesters, mostly activists from the movement Black Lives Matter (BLM), mainly used social media to organize and raise awareness, around 26 million people participated in the protests in the first month. Three days after the death, the hashtag #BlackLivesMatter was used almost 9 million times (Priniski et al. 2021) and anti-racist movements gained a lot of attention after the killing world-wide (Beaman and Fredette 2022). One example for such an activist tweet[10]:

> The racist violence that killed George Floyd, Ahmaud Arbery, and Breonna Taylor is not new in America. And what's captured on video represents only a fraction of the violence that Black Americans experience, some of it while in police custody. pic.twitter.com/X2SkYamKIQ — Elizabeth Warren (@ewarren) May 30, 2020

Even before the killing, there were several racist discussions throughout the COVID-19 pandemic, mainly from the right wing concerning the origin of the virus but also about the high death rates of ethnic minorities. However, these topics and tweets about racism in general grew a lot after the killing. Only 12% of the racism-related tweets mentioned George Floyd or Black Lives Matter (M. Thelwall and S. Thelwall 2021), indicating that racism was discussed on a broader level. Thus, not only the tweets against racism increased but also the ones spreading racism making it a very important topic in the society, in particular since these events.

2.3. Target languages

In this section the situation of hate speech and hate crimes is presented for each studied language with an exemplary country.

2.3.1. Hate in German/Germany

In Germany, there has been an increase in debates around (racial) hate speech[11] and its connection to physical violence since the refugee crisis in 2015 (Schröder 2020). Beforehand, racism was mostly marginally publicly debated[11] in the last decades with the exception of a few occasions like physical, racially motivated violence that were publicly discussed. Such violence has also happened in the last years. As the politician Walter Lübcke supported refugees in his political work, he has been the victim of a lot of (online) hate and was killed in 2019 due to his involvement (Schröder 2020). Another example is the attempt of a right-wing extremist to kill jews in Halle, Germany, in 2019[11]. There is a direct connection to online hate, too, as the

[10] Collection of tweets concerning the murder of George Floyd, Larry Ferlazzo: https://larryferlazzo.edublogs.org/2020/05/30/important-tweets-about-the-murder-of-george-floyd/

[11] DeZIM Research Notes Black Lives Matter in Europe, Noa Milman et. al.: https://www.dezim-institut.de/fileadmin/user_upload/Demo_FIS/publikation_pdf/FA-5265.pdf

perpetrator was radicalised and believed in conspiracy theories during his acitivities online. The racist assassination of Hanau followed this; eleven people were killed in 2020. A hate manifesto and video were published on social media by the assassinator (Schröder 2020).

One of the well-known far-right political parties in Germany is the AfD. It was founded in 2013 and has grown a lot due to opposing opinions to the welcoming refugee policy from then-chancellor Angela Merkel as well as incidents like the abundance of sexual assaults on New Year's Eve in December 2015[12] and the terrorist attacks on the satirical newspaper Charlie Hebdo[13]. The party expresses a lot of hate against immigrants (Fangen and Lichtenberg 2021). Their posts against refugees are correlated with physical attacks on them as these followed a peak in hate posts from the AfD[20].

Several anti-racist protests were held in the last years in Germany, including the Black Lives Matter protests following the murder of George Floyd, which received much attention and transferred the protest to the German context, too[11].

Legal situation The freedom of speech is protected in article 5 of the German constitution and protects the personal honor of the people (Vogel, Regev and Steinebach 2019). *Schmähkritik*, abusive criticism[14] for example is forbidden but depends highly on the context whereas incitement can lead to imprisonment of 5 years. The Netzwerkdurchsetzungsgesetz (NetzDG) from 2018 penalizes social media platforms when they do not delete reported content. It has been criticized as not being strong enough, the time limit not being realistic with not enough time to investigate the reported content and not covering all cases. The law affects only the biggest social media platforms and has raised data protection concerns when in 2020 it was decided that the important hate speech cases should be forwarded to the Bundeskriminalamt directly (Schröder 2020). Additionally, as the deadline of deleting content is tight, the platforms tend to delete more than necessary which endangers the protected freedom of speech (Vogel, Regev and Steinebach 2019).

2.3.2. Hate in French/France

In France, there is a strong conception of equality which is derived from the French Constitution (Article 1), stating that all citizen are equal without any distinction of race. The mention of race was removed in 2018, as it contradicted universalist principles. At the same time, it was criticized for denying racial violence and discrimination as well as limiting the consciousness of racial inequalities in the society[15] and supporting anti-racist movements (Goldman 2020). There are no statistics about racial profiling and almost no studies on migrant populations (Wang et al. 2021).

Despite the rare discussions on hate, there are a lot of hate crimes happening in France. The majority is based on racism or xenophobia (Hassan 2018). In general, the overall racism towards immigrants from former colonies and Islam has been increasing[15]. In 2020, there were 2672 hate

[12] Sexual assaults on New Year's Eve 2015

[13] Terrorist attacks on Charlie Hebdo

[14] The relationship of freedom of expression and protection of honour when making collective judgments about soldiers

[15] Race: A Never-Ending Taboo in France, Jean Beaman: `https://gjia.georgetown.edu/2021/04/01/race-a-never-ending-taboo-in-france/`

crimes reported by the police[16]. Among them, 58% were racially and xenophobically motivated. Online hate speech has been increasing, especially against migrant or Muslim communities (Vanetik and Mimoun 2022). Hate speech is rarely reported to the authorities though as those affected fear they are not taken seriously as they do not have a lot of confidence in the police[17].

During the outbreak of Covid-19 the discrimination against people perceived as *Chinese* has changed and increased as they were considered culpable because of the perceived poor hygiene standards in China. *Chinese* generalizes all Asians as both terms are often used with the same meaning in French and thus other minorities were discriminated against, too; especially when they began wearing masks before recommended by the French authorities (Wang et al. 2021).

Similarly to the US, People of Colour are affected by police violence in France. One specific example for police violence is the death of Adama Traoré in 2016. This has started protests similar to the ones from the BLM movement in the US protesting against racial discrimination for several weeks in Paris. However, the movement is not as strong as in the US. They started again after the killing of George Floyd (Goldman 2020). After two more racist violence crimes in 2020, the public debate around race increased more[15]. However, racial profiling and racism in general are still deemphasized in public debates by constructing a big contrast to the situation in the US (Beaman and Fredette 2022). Hence, racial inequality is perceived to be a rare exception in France and anti-racist movements are considered *un-French*, easing the spread of racism (Beaman and Fredette 2022).

Legal situation France prohibits hate speech, also on the internet (Chiril, Benamara et al. 2019). Since 27 January 2017, the penalties of discrimination have been doubled (Chiril, Benamara et al. 2019). In July 2019, France voted at a national assembly for a new law to fight hate speech online based on the German NetzDG (Maria Constantinou 2021) (see Section 2.3.1). However, the core of the law, requiring the platforms to take down a reported post in 24h, was striked upon. It would conflict with the freedom of speech as the pressure of the law would have led to censorship from the platforms[18].

2.3.3. Hate in Amharic/Ethiopia

The social media usage in Ethiopia has been influenced by a digital infrastructure that is not well established. At the same time, there are heavy regulations on the social media usage from the government. In 2019, the internet penetration was around 3%, the digital infrastructure development has been slow compared to other countries. The overall development of digital media was not enforced, but social media was controlled by the government (Demilie and Salau 2022) and critical voices using the internet were suppressed. When the Ethiopian People's Revolutionary Democratic Party (EPRDF) ruled, they tried to balance freedom of expression and using the media for their own goals (Pohjonen 2019). Websites contradicting its ideology were censored and bloggers or journalists raising critical voices were arrested (Pohjonen and Udupa 2017). Despite the challenges some online spaces already emerged in the beginning of the century (Pohjonen 2019) as an alternative to the censored press (Demilie and Salau 2022), co-occuring with first spreads of hate speech. This could especially be seen in debates

[16] Office for Democratic Institutions and Human Rights: https://hatecrime.osce.org/france

[17] Discours de haine et violence, Commission européenne contre le racisme et l'intolérance (ECRI): https://www.coe.int/fr/web/european-commission-against-racism-and-intolerance/hate-speech-and-violence

[18] French Court Strikes Down Most of Online Hate Speech Law: https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html

around the 2005 elections (Pohjonen 2019). Academic research on general online practices or communication has been sparse (Pohjonen and Udupa 2017).

After political reforms in 2018, the regulations changed and people could express their opinion more freely and gained more political liberties while still risking to be fined or imprisoned due to social media activities in the following years (Demilie and Salau 2022). The amount of tweets written in Amharic has increased (Yimam, Ayele and Biemann 2019), including hateful tweets with heavy consequences from the government such as turning off the internet for the whole country. The reforms eased academic research on hate speech but the resources are still sparse (Demilie and Salau 2022).

Legal situation The 2009 Anti-Terrorism Proclamation allowed the government to prohibit speech that was destabilizing the country with 15 years of imprisonment or death. Civil unrest further enforced criminalizing *negative speech* which was considered any kind of activity that "could create misunderstanding between people or unrest" (Pohjonen 2019). In 2020, a hate speech prevention and suppression proclamation was developed to combat hate speech and fake news. It has enabled more freedom for the citizen but at the same time it has been criticized to violate free speech and human rights as social media activities can still lead to imprisonment (Demilie and Salau 2022).

2.3.4. Hate in English/ the United States

The freedom of speech is considered very important in the United States of America[20]. In 2015, people were asked if it should be possible to offend minorities publicly. Compared to other regions of the world, people in the US agreed most often[20] (a median of 67% in the US compared to 35% globally). Since the presidential election of Donald Trump in 2016, the amount of hate speech and hate crimes has increased a lot (Vanetik and Mimoun 2022). Hence, in 2018, 37% of the adults in the US had experienced harassment online[20], 8% of them because of their race or ethnicity[19]. 85% of the Americans wanted improved resources for fighting against cyber hate[20] by the government. 79% find that the platforms have a responsibility to handle problematic content[19].

In the US, there have been several hate crimes related to racist activities on social media. Among them there is an extremist who shot on a synagogue in Pittsburgh in 2018. Being part of *Gab*, a platform famous for extremist content, he had beforehand spread racist conspiracy theories originating from the French far right[20].

Fighting against racial discrimination, especially profiling, the BLM movement in the US has developed a global infrastructure and grown a lot in seven years (Goldman 2020), in particular after the murder of George Floyd (see Section 2.2.1).

Legal situation The United States of America protect hate speech under the First Amendment of the US constitution as freedom of expression (Chiril, Benamara et al. 2019). Following the Telecommunications Act of 1996, the platforms cannot be hold liable for content from the users[20]. The way to handle problematic content therefore depends on the social media platforms whilst traditional media can be held liable for the content they publish[20]. In contrast to the

[19] Online Harassment 2017, Pew Research Center: `https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/`

[20] Hate Speech on Social Media: Global Comparisons, Council on Foreign Relations: `https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons`

EU, they only incite scrutinized violence but not denying genocides or hatred[20]. Exceptions to content-based regulation are libel, defamatory contents and incitement to violence (Siapera 2019).

2.4. Handling Hate Speech

There are several approaches to combat hate speech. One possibility is to directly react when witnessing or experiencing hate speech by interacting with the person affected, with the aggressor or reporting the behavior on the platform[21]. In the US, 65% of the witnesses react at least in one of these ways[21]. Initiatives like *Take back the tech*[22] try to help those affected by providing resources and strategies to handle it whereas other enforce counter speech expressed as a response to hate speech.

The platforms themselves try to filter the posts or to review and delete them when being reported as they are forced to by law in Germany. As humans alone could not stem the huge amount to be classified (Vanetik and Mimoun 2022), content moderation algorithms that detect hate speech in the posts are often used and combined with content reporting from the users and human content moderators. The contractors however suffer from the amount of content and the consequences from being exposed to hate. Since the rules which regulate the removal of content can be unclear and the perception of hate is subjective, they are usually not consistently applied by the staff[20]. This can make it more intransparent or frustrating for the users who reported the hate.

2.4.1. Automatic Detection

Hate speech detection can be simplified as a binary classification: hateful or not. Depending on the use case, offensive language or other types of harmful content can also be detected as well as specific kinds of hate like sexism or racism (Chiril, Benamara et al. 2019).

However, these human and machine classification systems are not yet well adapted to and equipped for all languages. For example, in 2015 Facebook only employed two fluent Burmese speakers even though the anti-muslim violence in Myanmar was already known to be a risk on the platform and thus, the amount of hate was not manageable by two humans[20].

Several classification systems exist for a specific languages such as Spanish or Italian (Chiril, Benamara et al. 2019). Low-ressource languages such as Amharic lack linguistic and natural language processing resources as well as datasets, thus making research in automatic hate detection more difficult (Demilie and Salau 2022). Some multilingual models already exist; for example one for English and Spanish (Chiril, Benamara et al. 2019) which has been trained on each language separately.

The models themselves still have difficulties recognising hate speech reliably and in all contexts, due to a number of reasons:

- Hate speech itself is hard to agree upon. Thus, when humans annotate a dataset, they do not always agree on the correct classification (see Section 2.1).

[21] Online harassment, digital abuse, and cyberstalking in America, Data & Society, Amanda Lenhart et. al.: `https://datasociety.net/library/online-harassment-digital-abuse-cyberstalking/`

[22] Take Back The Tech!: `https://takebackthetech.net/be-safe/hate-speech-strategies`

- Hate speech can be context specific or depend upon the community tolerance (ElSherief et al. 2018).
- Hate speech can be very subtle. It can be hidden in humour or the context needs to be known to detect it, see Table 2.1.
- The contexts on the internet change rapidly, and thus neologisms or new semantics occur that also need to be recognised by the models.
- People talking about discrimination in a neutral way use similar words which makes it harder to distinguish between facts and hate speech (Siapera 2019).
- Some marginalized groups use previously negatively connotated words to refer to themselves as an act of empowerment. As the systems are usually only trained on the negative meaning, a consequence is that statements from marginalized groups are rated more hateful than corresponding ones from non-marginalized groups (Mozafari, Farahbakhsh and Crespi 2020).
- In academic research, the use of very different corpora makes it difficult to evaluate and compare different classifiers (ElSherief et al. 2018).
- The studies define the problematic speech differently which makes it more difficult to compare the results (Schmidt and Wiegand 2017).
- Even though keyword lists are used, users may use abbreviations or intentional misspellings (ElSherief et al. 2018) as well as dialects or no punctuation (Yin and Zubiaga 2021).
- The annotation takes a lot of time and resources and the proportion of hate is small (ElSherief et al. 2018).
- The classification models have been trained on a specific context or platform and are not generalisable (Yin and Zubiaga 2021).
- Racism appears often implicitly in the form of stereotypes making it harder to be detected (Vanetik and Mimoun 2022).

Annotation

As the detection models need example classifications to be trained on, a common approach is to annotate existing posts by humans. They usually classify it as hate or not hate and this data can be used for supervised learning algorithms in order to detect hate speech. Examples for annotation strategies are given in Chapter 3.1 and an outline for the one used in this project is given in Chapter 4.1.

Conclusion

Hate speech has been defined together with offensive speech and harassment. Different forms and consequences of hate speech have been outlined. Racism has been defined and the example of the killing of George Floyd shows a connection between physical hate crimes and online hate speech on a global extent. One country has been presented for each language as an example. The respective situations concerning hate speech, hate crimes and discussion around racism

have been outlined. Other countries, in which the same languages are spoken will need to be taken into account in future research. The problem hate speech and its relation to physical violence exists in all illustrated countries. The ways the governments sanction hate speech differ and also the way that racism is treated. Thus, it is important to treat the hate speech detection as a global problem but take into account local differences. Especially low-resource languages still lack research and resources to study them. In the next chapter I will introduce related work to multilingual hate speech detection.

3. Related work

Here I include an overview of related work on multilingual and racial hate speech detection. The work will be presented together with their results and limitations. Additionally, several studies on improving the annotation process are presented.

3.1. General hate speech detection

3.1.1. Classifiers

In order to detect hate speech automatically, several classification models have been build specialising in languages, domains, platforms or other features. Some of these classification models for detecting hate speech are presented in the following sections.

The language model family Bidirectional Encoder Representations from Transformers (BERT) is the current state-of-the-art (since 2019) for classification models (Yin and Zubiaga 2021). The BERT-models can easily be adapted and fine-tuned to specific tasks. For example, they were used to determine if a group identifier like "gay" was used in an offensive way (Kennedy, Jin et al. 2020). Post-hoc explanations from fine-tuned BERT classifiers were used and regularization methods were used later on to also include the context for the model learning. They could reduce false positives for out-of-domain data.

Focusing on race, nationality and religion, the group from Njagi et al. (2015) created a classifier detecting hate speech for these target categories. The classifier uses subjectivity detection – a sentiment analysis – for ranking the polarity of an expression. It is build upon a lexicon based on subjectivity and semantic features.

Several experiments have been done on a dataset collected from the white supremacy forum Stormfront, differing hate speech and not hate speech. During the annotation it was possible to access the context of a post, reducing the bias (Gibert et al. 2018).

The majority of the studies used Twitter data (Mathew et al. 2021), (Vidgen and Derczynski 2020). Many of the published detection models still have limitations. On a survey from Yin and Zubiaga (2021), they found that most detection models' performances are overestimated and the generalization is poor. However, using data that is not platform or domain specific helps generalising the model.

Salminen, Hopf et al. (2020) already considered multiple social media platforms. They used data from YouTube, Reddit, Wikipedia, and Twitter to build several classifiers. Their best model using XGBoost was considered generalizable, it achieved a F1 score, a measurement of statistical methods ranging from 0 to 1 which is perfect) of 0.92.

3.1.2. Data collection

There are different approaches used to collect data for building a model. A large collection of datasets exists for English. For German and French, there are fewer ones (see Figure 3.1).

As the amount of hate speech in a document is usually very sparse, it is difficult to develop a representative corpus which is generalisable to other kinds of hate speech. Using strategies like keywords or topics for filtering can thus be useful but can also introduce biases at the same time (Kennedy, Atari et al. 2020).

Kennedy, Atari et al. (2020) have collected around 28,000 posts from *Gab*, a social network and build a corpus[1] which is one of the largest corpora for hate speech detection. They have annotated each post by at least three annotators that were trained beforehand. This procedure is very common for corpus building.

Alternatively, one can also collect data from a specific geographical region. In this case, posts in Ireland related to extremism were collected with the aim of distinguishing acceptable race talk and racist utterings (Siapera 2019).

Another approach is to base the research theoretically on critical race theory and extract criteria to use them for annotating an already published corpus (Waseem and Hovy 2016). A linguistic analysis was done and a dictionary containing the most indicative words of their research was published, too.

Hatespeech based on social biases can be uttered implicitly. The work from Sap, Gabriel et al. (2020) has formalised this phenomena and build a corpus[2] with 34000 implications of stereotypes and biases from social media posts.Another form of problematic speech is harrassment. A corpus specifically for harassment research including 35k expert-annotated tweets was published (Golbeck et al. 2017).

To learn a representation of hate and to ease the annotation (Rizoiu et al. 2019), methods were developed to automatically analyse a text starting from small, unrelated datasets. The combination of a deep neural network with transfer learning has a prediction correctness is the macro-averaged F1 of 78% and 72% in the first (detect racist and sexist text) and second (detect hate and offensive text) task, respectively. The output is called a Map of Hate and should be a human-interpretable way to visualize the type of hate entailed in a text by word and sentence embeddings.

Based on the analysis and categorisation of microaggressions, which are subtle biases, two datasets[3] were build to further work on microagreession annotation (Breitfeller et al. 2019).

Lexical methods

Often, lexical methods are used to retrieve social media posts based on the entries in a lexicon and thus to build a dataset. In a study, Davidson, Warmsley et al. (2017) analysed the quality of lexical methods. They proved to be more effective to detect offensive language but not for

[1] Corpus with posts from Gab: `https://osf.io/edua3/`

[2] Data from `https://homes.cs.washington.edu/~msap/social-bias-frames/SBIC.v2.tgz`

[3] Datasets to analyse microaggressions: `https://drive.google.com/drive/folders/1bKf8PQuuOk7z3ehgAcmTLjmK5Cb86ZTz`

hate speech when comparing the results of the annotation with their dataset[4]. They also found that it depends on the type of hate how they are classified: racism and homophobia tend to be hate more and sexism is more offensive. It is more difficult to classify hate when there are no keywords (Davidson, Warmsley et al. 2017).

In order to build a dataset to detect racism in Dutch Social Media, a dictionary-based approach was chosen in another study (Tulkens et al. 2016). Three different dictionaries were selected, one with training data and the others were an augmentation by retrieving racist and neutral terms or adding general words. This set with the removed incorrect expansions performed the best with an F-score of 0.46 for racist comments of the test data.

Another approach for detecting offensive words was to only use a swearword lexicon (Klenner 2018). It was build by retrieving the 300 most frequent words of Facebook post from a German right party and added them to an existing one. They were classified as positive or negative for the categories emotion, moral or appreciation.

One way to collect and filter tweets was to collect all that contained at least one offensive word. As this is not a reliable source for an offensive meaning, human annotators reduced this uncertainty (Rezvan et al. 2018) by annotating the 24,189 tweets in the dataset containing including racial harassment[5].

There are less annotated datasets that deal with racist speech than for general hatespeech, in particular for French language (Vanetik and Mimoun 2022).

3.2. Languages in hate speech detection

In this section, hate speech detection for the languages German, French and Amharic are presented. The majority of works on hate speech detection consider only English content (Vanetik and Mimoun 2022) and thus, most of the datasets for hate speech annotation use English posts (Yin and Zubiaga 2021). English-speaking countries such as the United States, the United Kingdom or Australia are among the highest-ranked countries for the amount of publications on online hate speech research and have influenced the research on machine learning and text classification algorithms for hate speech detection. Germany is represented in the top 7 of the analysed topics on online hate research in the last thirty years. France and Ethiopia on the other hand are not mentioned in the survey, indicating that their research has not grown as much (Tontodimamma et al. 2021).

3.2.1. Multilingual classification

As not everything on social media is published in English, research on other languages and multi-lingual settings is required and already evolving. For example, Turkish, Danish and Slovene were already analysed to detect offensive speech among others (Vanetik and Mimoun 2022). A shared task was published to detect hate speech both in a multilingual and domain-focused environment, more specifically for English and Spanish and targeting immigrants and women[6].

[4] Dataset from Automated Hate Speech Detection and the Problem of Offensive Language: `https://github.com/mayelsherif/hate_speech_icwsm18`

[5] Harassment corpus: `https://github.com/Mrezvan94/Harassment-Corpus`

[6] SemEval 2019 Task 5 - Shared Task on Multilingual Detection of Hate: `https://competitions.codalab.org/competitions/19935`

However, the distribution of languages of already published datasets does not reflect the spoken language distribution in the world (Vidgen and Derczynski 2020).

Several experiments for multilingual hate speech detection were conducted in one study (Aluru et al. 2020). In nine languages they have tested the models and found that for low resource languages a LASER embedding with logistic regression performs best whilst BERT works good for high-ressource languages. For zero shot training, Italian and Portuguese work well.

XHATE-999 for example is a dataset for multiple domains and languages which can be used for evaluating classifiers (Glavaš, Karan and Vulić 2020). It has been used in a zero-shot transfer learning; the respective language was modeled on an abusive corpus and the domain was adapted. Ousidhoum et al. (2019a) have developed a multilingual dataset to evaluate different hate speech detection approaches.

A specific neural network architecture in a multilingual setting is proposed in (Corazza et al. 2020). It is tested and analysed to better understand the influences of the components on datasets in the three languages English, Italian and German. There is also a supervised approach for English and French considering on problematic speech against immigrants (only in English) and women as target groups including feature-engineering and neural approaches (Chiril, Benamara et al. 2019).

The Transformer language models (Raha et al. 2021) for hateful/offensive/profane texts in English, German and Hindi have achieved F1 scores of 90.29, 81.87 and 75.40. The model is based on a pre-trained text encoder which is also Transformer-based. By investigating Transformer language models in a multilingual setting in another study (Roy et al. 2021), hate speech could be classified in English, German and Hindi.

Low-resource hate speech detection

As there are not a lot of resources for some languages, zero-shot and transfer learning methods are often used. One example for a zero-shot hate speech classification was done in Urdu (Khan, Shahzad and Malik 2021). Logistic regression and deep learning performed best (F1 score of 0.906 for distinguishing between Neutral-Hostile tweets, and 0.756 for distinguishing between Offensive-Hate speech tweets). They have used 5000 Roman Urdu tweets for building a problematic speech corpus and annotated them based on three types: Neutral-Hostile, Simple-Complex, and Offensive-Hate speech.

Another transfer learning approach is to use data from high-resource languages with a Convolution Neural Network to identify intents with a character probability map. They achieved significant results for Sinhala and Tamil (Karunanayake, Thayasivam and Ranathunga 2019).

An annotated Greek dataset has been developed and evaluated with 4,779 tweets, either offensive or not offensive (Pitenis, Zampieri and Ranasinghe 2020): Detecting offensive language in Greek, including a tweet dataset.

3.2.2. German content

There are already several studies that detect German content; some of them are presented here. The paper (Vogel, Regev and Steinebach 2019) analysed German content in social media which is radical. They used the k-nearest neighbours method to detect hate speech in the tweets.

An accuracy of 82% was achieved for their dataset. They emphasize using various domains to reduce biases in the writing style or on a certain topic.

The GermEval 2018 Shared Task put emphasis on identifying offensive language, more specifically to classify German tweets in two steps: a coarse and a fine-grained classification (Klenner 2018). Another study (Bretschneider 2017) worked in offensive language directed against foreigners. They also developed a dataset containing the statement and its target to improve the automatic detection. A severity value was used to also detect hostility.

Jaki and Smedt (2019) analysed hate speech on Twitter from right-wing people linguistically. They showed that only a fraction of tweets that are perceived hateful are illegal; the rest is protected by the freedom of speech. However, there is no study that detects racial hate speech in German in a multilingual approach.

3.2.3. French content

Hate speech, racism and racial profiling are less studied in French than in English, respectively (Vanetik and Mimoun 2022). Chiril, Moriceau et al. (2020) created a French corpus of sexist tweets by a keyword list, being the first to detect sexism and multitarget hate speech in French. Their best classifiers achieved 0.788 accuracy and 0.780 F1 for hate speech detection and 0.822 accuracy and 0.688 F1 for sexism detection.

There are only two french datasets listed in *The Hate Speech Dataset Catalogue*, a collection of datasets for hate speech research. One of them is the COunter NArratives through Nichesourcing (CONAN) (Chung et al. 2019) with counter-narratives (an informed textual response) by experts from NGOs in a multilingual approach. It includes content in French, Italian and English and is limited to Islamophobic content. The counter-narratives are considered as an alternative to deleting content or blocking users. Additionally, they include information on the expert demographics, types of hate and responses as well as data augmentation like translation. The other entry is the Multilingual and Multi-Aspect Hate Speech Analysis dataset (MLMA) containing 4014 hate speech comments with various multi-class labels.

There is one recent study that analysed French racist tweets by building a dataset French Twitter Racist speech dataset (FTR) collected from the twitter stream based on a list of racist terms. They were then manually annotated, distinguishing racism and no racism but not hate or offensive. Vanetik and Mimoun (2022) achieved similar accuracies for tf-idf and BERT sentence embeddings. Pre-trained fine-tuned BERT models resulted in lower scores and is explained by the difficulty of detecting racism compared to detecting hate speech and the size of the dataset (Vanetik and Mimoun 2022). They also found out that transfer cross-lingual learning did not work for racist speech in French. The mixed domain training with a general hate speech and their specific racism data set improved the results (Vanetik and Mimoun 2022).

There is little research on racist hate speech detection and the one study found only considers racism, not the distinction of hate speech and offensiveness.

3.2.4. Amharic content

The hate speech research has started to evolve in recent years. A first dataset of abusive Twitter data was collected (Yimam, Ayele and Biemann 2019) based on keywords for hate and offensive speech.

Language	Dataset	Literature
ENG	HS: Twitter (MLMA)	Ousidhoum et al. (2019b)
ENG	HS: Yahoo!, American Jewish Congress	Warner and Hirschberg (2012)
ENG	HS: Twitter	Waseem and Hovy (2016)
ENG	User comments from Fox news	Gao and Huang (2017)
ENG	Offensive tweets with targets	Zampieri et al. (2019)
ENG	Racist and sexist tweets	Waseem (2016), Waseem and Hovy (2016)
ENG	HS: Twitter (hateval)	Basile et al. (2019)
ENG	Abusive tweets	Founta et al. (2018)
EN, GER	Offensive, hateful, profane tweets (hasoc)	**hasoc**
GER	Refugee crisis related hate tweets	Ross et al. (2016)
GER	Offensive Facebook posts against foreigners	Bretschneider (2017)
GER	Offensive tweets	Bai et al. (2018)
FR	HS tweets (MLMA)	Ousidhoum et al. (2019b)
FR	Sexist tweets	Chiril, Moriceau et al. (2020) Chiril, Benamara et al. (2019)

Table 3.1.: Overview of different hate speech (HS) and offensive speech datasets

By using deep learning methods, there were multiple studies to detect fake news as well as projects for computational linguistics in Amharic. By using word embeddings, an accuracy of 99.36% could be achieved to detect Amharic fake news. Another study collected data from facebook and used two expert annotators on a balanced dataset (Demilie and Salau 2022). A resulting Convolutional neural network (CNN) achieved an accuracy of 93.92% and an f1-score of 94%. However, the resources with manual labels remain few (Demilie and Salau 2022).

3.2.5. Published Datasets

A selection of already published datasets for hate speech detection are shown in Table 3.1. The links to access the corresponding datasets are published in the Github repository[7].

3.3. Annotation and Biases

In order to create datasets for the detection models, human-annotated social media posts are the most common approach to collect data. This method has the advantage of being easy to implement and having a large number of people (Vidgen and Derczynski 2020) However, classifying a text as hate can be very subjective and thus it is difficult to create a reliable dataset (Ross et al. 2016). Furthermore, there are more biases such as the annotator's knowledge which has been compared in (Waseem 2016). Experts perfom better than amateurs who label a text more likely as hate speech.

[7] https://github.com/bickbeermoos/multiling_hatespeech/blob/main/Dataset_links.md

The generalisability of hate speech detection models was analysed (Yin and Zubiaga 2021). It is found to not be very strong due to a number of reasons. Both in the datasets and in the models arise problems: the way the dataset was created, general limitations of Natural Language Processing (NLP) as well as the variety of online hate speech and their connections. As there are often small datasets, overfitting is a common problem and biases from the datasets are forwarded to the models. These biases are either methodologically or from the society (Yin and Zubiaga 2021). Racial biases for example can arise from oversampling specific keywords (Vidgen and Derczynski 2020).

An example for biases in the datasets is the discrimination against African-American English (AAE). As Davidson, Bhattacharya and Weber (2019)found out, when training classifiers on annotated datasets for hate speech detection, the prediction for tweets in African-American English differs from those in Standard American English. The African-American ones are classified as abusive more often, indicating a systematic bias discriminating those that should be protected. This bias was also found by Mozafari, Farahbakhsh and Crespi (2020) who used transfer learning on a pre-trained BERT model to detect racism, sexism and offensive and hateful tweets on annotated datasets. When analysing their classification, they could also see the negative bias towards AAE.

The bias could already reduced by regularization (Mozafari, Farahbakhsh and Crespi 2020) and making the annotators aware of possible race primes (Sap, Card et al. 2019). As there also may be imbalance between the occurrences of words in general and in those classes considered hateful or toxic, Dixon et al. (2018) have developed a method to add data to reduce unintended biases in the data.

Another bias derives from the content creators as for example in the popular dataset[8] from Waseem and Hovy 70% of the sexist and 99% of the racist content were published by 2 or respectively 1 person(s) (Vidgen and Derczynski 2020).

3.4. Comparative studies

Few studies have compared hate speech in different countries (Pohjonen 2019) have dealt with Finland and Ethiopia whilst (Pohjonen and Udupa 2017) compared India and Ethiopia.
Udapa et al. analyses in an ongoing study[9] the differences of extreme online speech in several countries. In another study, asked annotators from 50 different countries were asked to annotate the same tasks and could observe several differences between the countries but even more on a subject level (Salminen, Veronesi et al. 2018).

Conclusion

Related work has been presented for building a corpus by lexica and for building classifiers. The state-of-the-art language model family BERT was introduced. Studies on multilingual and those who focus on French, German or Amharic were described. There are already several multilingual studies and datasets but they do not fit the need of my research questions which is why my

[8] Data for Are You a Racist or Am I Seeing Things? and Hateful Symbols or Hateful People?: `https://github.com/ZeerakW/hatespeech`

[9] `https://www.research-in-bavaria.de/smart-village/media-anthropology`

work is presented in the next chapters. More specifically, there are very few comparative studies and none could be found that explores the cultural and linguistical specificities of hate speech in different languages.

4. Data collection

In this chapter I will explain the process of collecting data with which the detection model should be trained. I will describe how the tweets were filtered and preprocessed such that they could be annotated. The collected datasets are analysed and the annotators' demographics are presented.

4.1. Data collection strategy

There are already several datasets on hate speech detection as outlined in Table 3.1. However, they do not suffice this thesis' purposes as they do not consider racial hate speech in multilingual settings. For this thesis project, three datasets were build: a German, a French and an Amharic one.

The data was collected from Twitter as the platform is one of the biggest platforms having to deal with hate speech and the tweets are easily accessible for research purposes. The preprocessing of the tweets was similar to the one for the model HateXplain (Mathew et al. 2021).

In order to increase the amount of possible hate and offensive speech, a keyword list was build for each language. Despite the possible new biases, it is especially useful to filter offensive speech as described in Section 3.1.2.

When selecting the data based on a target group, only the dimension of racism has been considered which aligns with the findings from Vidgen and Derczynski (2020): due to the polarized society, it can make sense to focus on a specific context and to not build a general model. More specifically, only tweets that were published in the month after the killing of George Floyd[1] were selected for this thesis. This is due to the fact that this was a global acceleration point for racism and debates around it.

Hence, for this thesis, Twitter data containing potential racial hate speech and offenses was collected. The tweets were then annotated in order to build machine learning models, which are constructed to recognize racial hate speech in tweets in multiple languages, a variety of low- and high-resource languages. In order to collect the data for the models, the following strategy was chosen:

1. Collect and concatenate keyword lists (see Table 4.1.1)
2. Filter tweets:
    a) Access the 1% Twitter dataset
    b) Use only tweets published between 20th May-20th June 2020 (the month after the death of George Floyd)
    c) Remove truncated tweets
    d) Remove retweets
    e) Remove duplicates

[1] The New York Times: How George Floyd Died, and What Happened Next

f) Detect language with pycld2
g) Detect at least one language
h) Detect at least one keyword
i) Keep the ones where French/German/Amharic was detected, separate them accordingly and keep 5000 tweets per language

3. Annotate remaining tweets with Toloka[2], for the detailed approach see Section 4.2

Following these steps I could obtain a French corpus and a German one with each around 5000 annotated tweets. For more details see Section 4.2.4. The Amharic dataset was selected from the same time span and preprocessed like in the paper (Yimam, Ayele and Biemann 2019). It contains 2000 tweets. As outlined in the Chapter 3.1, there are already a lot of publications analysing English hate speech. Due to the limited scope of this thesis, there was no new English dataset created.

4.1.1. Lexica

I have compiled keyword lists that were used in other studies and searched for lists containing swear words to better detect slurs. These have been concatenated and used for filtering the tweets. As it can be seen on Table 4.1, the *Racial/All* column indicates how many keywords were marked as being a racial word in each list.

After concatenating, the French lexicon contained 3473 keywords and the German one 17367. The Amharic keyword list (Yimam, Ayele and Biemann 2019) contains 99 hate and 48 offensive keywords and has been developed with ten law, linguistic and social science experts. The most common hateful keywords are, among them two opposition parties:

- TPL (Tigray People's Liberation Front)
- OL (Oromo Liberation Front)
- enemy
- racist

The most common offensive keywords are:

- farmer
- dog
- donkey
- stench

It is already clear from these lists that the hateful keywords refer to politics and have very strong meanings. The offensive words however mostly refer to everyday life things.

4.1.2. Filtering the tweets

Tweets from the 1% Twitter stream have been downloaded and filtered based on keyword lists, similarly to the structure from ElSherief et al. (2018). There are 91 files with tweets published betweent the 20th May 2020 and the 20th Jun 2020 (the month after the killing of George

[2]Toloka Crowdsourcing: `https://toloka.yandex.com/`

Language	Name	Racial/All	Short title in the code
ENG	Profane words in different contexts	161/725	harass
ENG	Abusive words	0/8478	abusive
ENG	Bad words	0/1383	badwords
ENG	Hate speech lexicon of n-grams based on hatebase	0/178	hatebase_refined
ENG	Offensive, aggressive, and hateful words (hurtlex)	0/8228	hurtlex
GER	Offensive, aggressive, and hateful words (hurtlex)	0/ 5039	hurtlex
ENG	Common slurs, controversial topics, insulting patterns during debates (MLMA)	0/84	mlma_words
ENG	Expressions against immigrants	27/27	hateval
ENG	Racial hashtags (Basile et al. 2019)	03/03	ssrn
ENG	Racial slurs	0/ 2688	slur
ENG	Hateful keywords based on hatebase	0/51	icwsm
GER	Ethnoplausisms	40/40	wikipedia
GER	Ethnic slurs	119/119	wikipedia
GER	Insults	0/2180	insult
GER	Offensive words	0/11322	hyperhero
GER	Racial slurs	19/21	uni-graz
GER	Offensive words	19/19	rp
GER	Racial slurs	8/8	neuemedien
GER	Racial words	2/2	wireltern
GER	Offenses	0/315	sprachnudel
FR	Common slurs, controversial topics, insulting patterns during debates (MLMA)	0/70	mlma_words
FR	Hurtlex	0/5024	hurtlex
FR	French sexist words	0/156	sexist
FR	French swear words	0/17	iceberg
AM	Hateful and offensive keywords (Yimam, Ayele and Biemann 2019)	13/72	amharic

Table 4.1.: Keyword lists used for filtering the tweets

Floyd, see Section 2.2.1) with a total amount of 196 679 547 tweets. They were then further preprocessed, for example corrupt entries were removed. This leaves around 53 million tweets. They are then filtered based on the `created_at` date, deleted tweets are not taken into account and the language is recognized with `Pycld2`[3]. If the only language recognised in the tweet is *unknown* or if there are several ones recognised (and thus indicating code-switching), the tweet is removed. This is due to the fact that we did not want to require the annotators to know any other than the studied language. Usernames and links are replaced by @link and @username respectively for anonymization. Retweets are also deleted as they make it more difficult to understand the tweet as the context is missing. Abbreviations like *mdr* (mort de rire, laugh out loud) were kept as removing them would reduce the accuracy (Vanetik and Mimoun 2022). Tweets are considered if at least one word, which is not a stop word but which is contained in the keyword list described in Section 4.1.1, is included in the tweet. The corresponding code is published on github[4].

The Amharic data collection process was very similar, the tweets were filtered by their publication date. Retweets and near-duplicates were removed, only tweets that contain the keywords were selected (Yimam, Ayele and Biemann 2019).

4.2. Annotation

The annotation is done by crowdsource annotators as it is common in this research field (Vidgen and Derczynski 2020).

The crowdsourcing platform Toloka[5] was used to conduct the studies. Several pilot studies have been conducted for the respective languages before the main studies. For example, the Amharic pilot studies contained 200 tweets respectively and were used to identify and ban malicious annotators who responded arbitrarily.

There are different options, annotators can decide if a tweet contains hate speech or offensive content and if so, if it is racial (see Figure 4.1). There is also the possibility to select *Unsure*, giving the users the opportunity to indicate that a tweet is very hard to classify. This is also a basis for further research into hate speech detection.

The annotators were shown definitions and training examples beforehand as Ross et al. (2016) found that a given definition raises alignment of the opinion of the annotators. More specifically, a guideline (see Section A) was shown to the annotators and they had to successfully complete training tasks to qualify for the main annotation. An overview of the French study in Toloka is shown in Figure 4.2.

4.2.1. Selecting the users

For the French dataset, the crowdworkers need to have passed the French language test (see Figure 4.3 for the French language test) and live in France or Belgium. For the German one the country filter was first restricted to Germany and Austria but as there were too few crowdworkers, this restriction was removed during the annotation.

[3]Pycld2: https://pypi.org/project/pycld2/
[4]Github repository, Skadi Dinter: `https://github.com/bickbeermoos/multiling_hatespeech`
[5]Toloka Crowdsourcing: `https://toloka.yandex.com/`

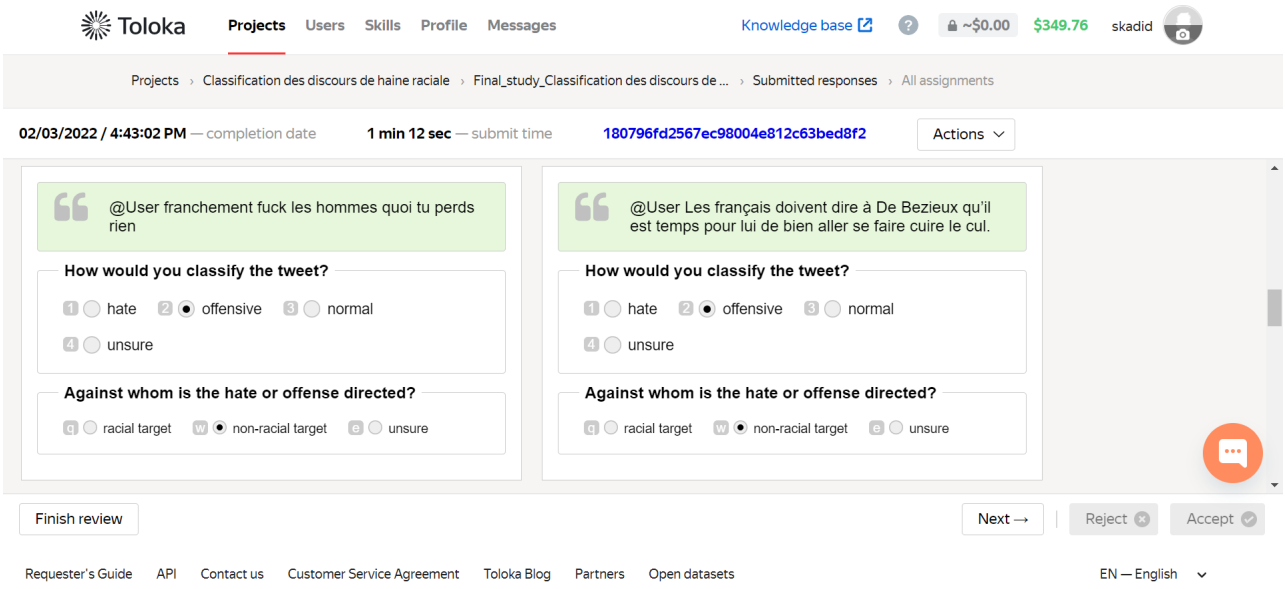

Figure 4.1.: Example of the French Annotation tasks

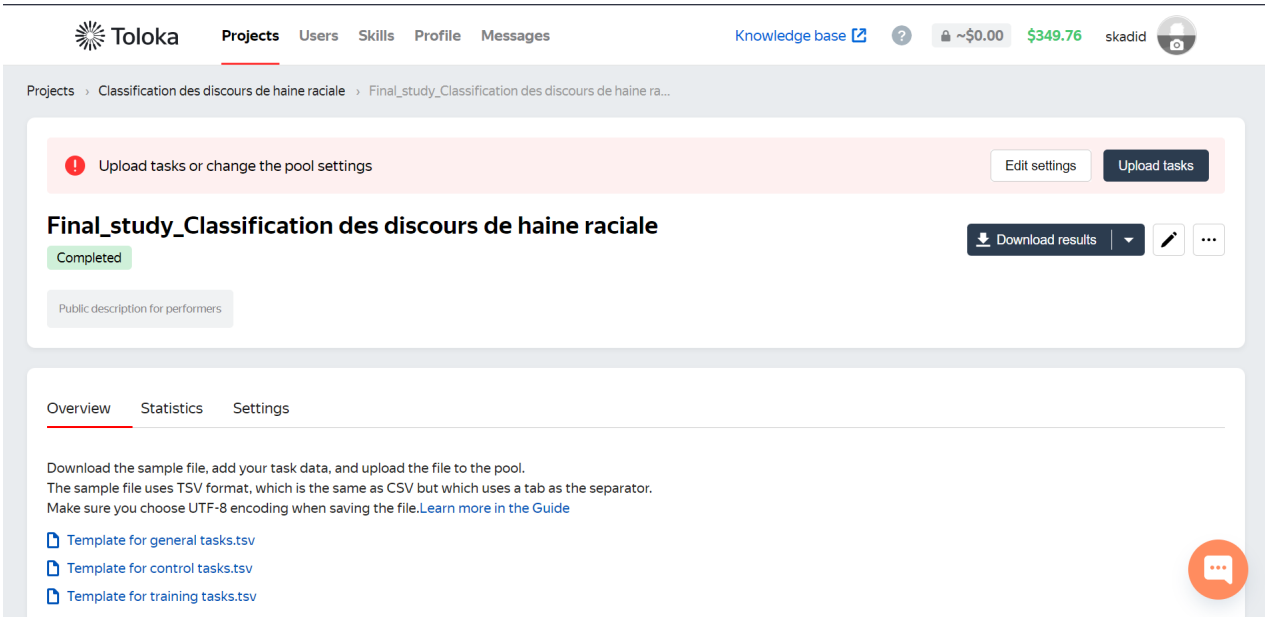

Figure 4.2.: The completed study of the French tasks

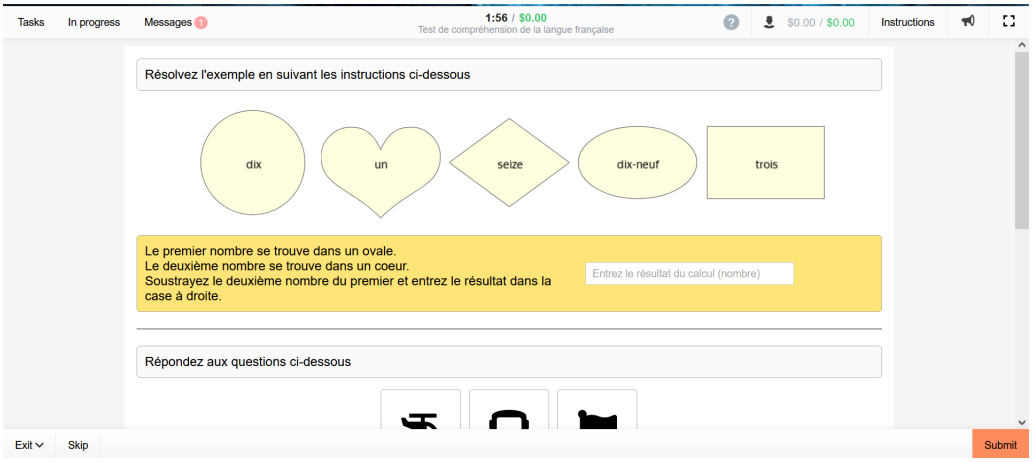

Figure 4.3.: The language test for French

The user also needed good reputation scores, i.e. being among the top 90% of the users. They have to pass a so-called training pool including two tasks structured the same way as the actual tasks to gain access to the proper task pool. This was to introduce them to the tasks and make sure they understand them.

4.2.2. Guideline

The annotation guideline used for annotation was inspired by Zampieri et al. (2019) and translated respectively for the different annotation languages. It can be found in the appendix in Section A. These are instructions for the crowdworkers to help them classify the tweets and setting the rules for the annotation. For example, they are asked to respect the privacy of the content creators as was done by Sap, Gabriel et al. (2020).

4.2.3. Risk mitigation

Risks that might arise for the annotators were mitigated as much as possible. One of them is acute stress or other mental health issues due to the problematic content of the tweets. The amount of tasks to annotate has been limited to avoid acute stress as well providing a crisis management resource[6] (only available in English) (Sap, Gabriel et al. 2020).

4.2.4. Annotation results

Each tweet was annotated by three annotators, the final classification was evaluated from these three annotations. The results, together with the country and the age of the annotators, could be accessed directly. Furthermore, a Dawid-Skene aggregation was conducted to get one response[7] for each tweet with a confidence. This is an aggregation method taking into account the response popularity and error matrices. As a result, for each tweet a classification together with a confidence score was obtained. The Table 4.2 shows a comparison of the annotated

[6] Crisis text line: `https://www.crisistextline.org/`

[7] Automatic Dawid-Skene aggregation: `https://toloka.ai/docs/guide/concepts/result-aggregation.html`

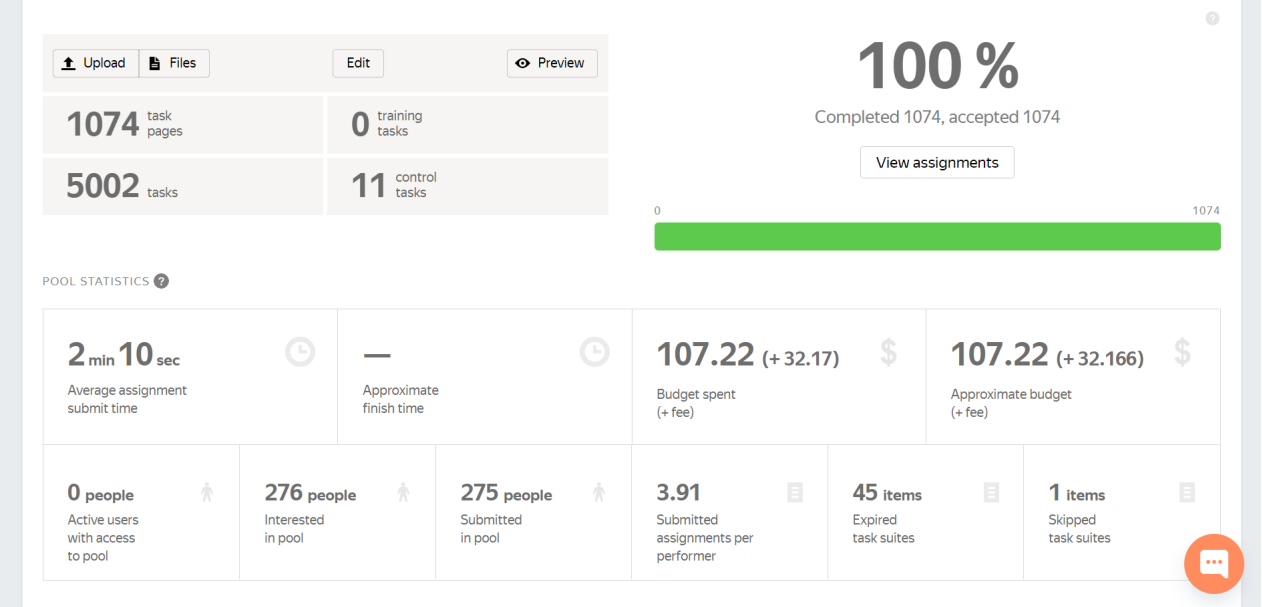

Figure 4.4.: Overview of the completed French annotation in Toloka

datasets in the three languages. For French and German, 50 random tweets were self-annotated by myself and used as a ground truth to measure the accuracy and F1-score of the annotations.

German tweets

There were 4999 German tweets annotated by 306 annotators. For the annotation, 86.88% of the tweets were considered normal, 8.78% hateful, 3.63% offensive and 6.6% are ties and 0.68% are unsure (see Figure 4.5). Out of those hateful, 80.44% are against a racial target, contrary to only 58.58% for the offensive comments (Figure 4.6). This can be explained by the fact that hate is often directed against a target group, but an offensive post is not necessarily directed against one. The amount of ties is almost a third for the offensive ones contrary to 13.45% for those hateful ones. These classifications were done with a majority voting, if at least two annotators agreed on a classification, this was done. A tie means that all three annotators selected different options.

The inter-annotator agreement is very low, the Fleiss Kappa[8] is 0.081 indicating a slight agreement.The Fleiss Kappa is a measurement to find the nominal scale agreement between raters, it is the generalization of the Cohens Kappa for more than two raters (Fleiss 1971).

The low numbers can be explained by the non-binary classification options and are common for abusive content annotation (Vidgen and Derczynski 2020). This finding indicates challenges regarding the research question which is further discussed in Chapter 6.1. Other reasons for this score might be:

- The annotation task was very difficult: defining hate is very subjective even when there are guidelines
- Some context is missing even though the tweets with lacking context were removed

[8]Inter-annotator agreement: `https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3`

	French	German	Amharic
Fleiss Kappa	0.3	0.0805	0.142
Amount of annotated tweets	5002	4999	1400
Amount of annotators	275	306	100
Mean age in years	31.11	32.97	25.38
Country distribution	265 FR, 8 BE, 3 other	43 DE, 6 AT, 257 other	66 ET, 19 other
Accuracy for 50 rand tweets	0.24	0.06	–
F1 score for 50 rand tweets	0.24	0.06	–
Racial accuracy for 50 rand tweets	0.12	0.08	–
Average time for 15 tweets	2 min 10 sec	3 min 14 sec	4 min 10 sec
Collected keywords	3473	17367	147

Table 4.2.: Comparison of the annotated datasets

- The training pool was maybe not sufficient to qualify the annotators
- The language test was not appropriate to check the language abilities sufficiently
- The guidelines might not have been considered or were too long/short/unclear
- Crowdsourcing tasks are not very popular in Germany (8% frequent crowdsourcers, (Pesole et al. 2018)) making it more difficult to find qualified crowdsourcers
- The payment might have been not sufficient to find qualified annotators
- Due to the different backgrounds of the annotators, they might classify differently as they are not equally sensitive
- They might lack knowledge to understand the meaning of a tweet (when it is in a slang or background information is important)

As the inter-annotator agreement was not high enough, the dataset was not used to build the model. Thus, the collected dataset is not used as a ground truth but rather a starting point for future research on agreements in German labeling. To further explore the different agreements, some examples are displayed in Table 4.2.4.

The complete agreements are as follows: 2429 times all three annotators selected normal, 57 times they selected hate and 32 times offensive. Not once were all unsure. For the majority, where two annotators agree: 1616 times normal, 352 times hate, 137 times offensive, 33 times unsure.

Ties There are 343 tweets that were classified three times differently. One example for a tie is the tweet:

Language	Classifications	Example	Translation	Labels
French	False predictions	@User avoue c'est toi qui cause cette phobie aux gens	@User admit it's you who causes this phobia to people	1*N, 2*O, 1*nrac, 1*U rac, 1*none: **tie**; self annotation: O
German	False predictions	@User Euch braucht kein Mensch mehr!! Ihr seid längst digital überholt!! Schaltet Euch ab. Ihr kostet nur unser Geld!!	@User No one needs you anymore!! You are long since digitally obsolete!! Shut down. You only cost our money!!	1*N,2* H,2*rac, 1*none: **H, rac**; self annotation: O nrac
French	correct predictions	@User @User @User Je vais te briser tes os je vais boire ton sang	@User @User @User I will break your bones I will drink your blood	3*H, 3*rac: **H, rac**; self annotation: H, rac
German	correct predictions	@User @User @User @User Hör auf dumm zu sein	@User @User @User Stop being stupid	3*O, 3*nrac: **O, nrac**; self annotation: O, nrac
French	complete agreement	Damon et Chloé j 'vous aime trop fort	Damon and Chloe I love you too much	N
German	complete agreement	@User @Link Eh. Da stimme ich ja zu. Aber das als große Reform zu verkaufen halte ich für Sinnlos.	@User @Link Eh. I agree with that. But selling this as a major reform I think is pointless.	N
French	majority voting	une ministre des sports qui lâchent ce genre de déclaration ? on se demande si on est vraiment en france	a minister of sports who drops this kind of statement? one wonders if we are really in France	1*N, 2*O,2*nrac, 1*none: O
German	majority voting	@User Dieses bitch Damn Klatsch ich dein Vater was los fühl ich	@User This bitch Damn gossip I feel your father what's going on I	1*N, 2* H; 2*rac, 1* none: H
French	tie	une excuse ça va pas suffir il faut apprendre à s 'éduquer et arrêter de faire du slut shame h24	an excuse won't be enough, you have to learn to educate yourself and stop slut shaming people 24 hours a day	1*N, 1*O, 1*U,1*nrac, 2*none
German	tie	@User @User @User Schon der grosse Philosoph Helge Schneider sagte dereinst: es gibt Reis	@User @User @User Already the great philosopher Helge Schneider said once: there is rice	N, H nrac, H rac

Table 4.3.: Examples for different agreements: normal (N), offensive (O), hate (H), unsure (U), racial (rac), non-racial (nrac)

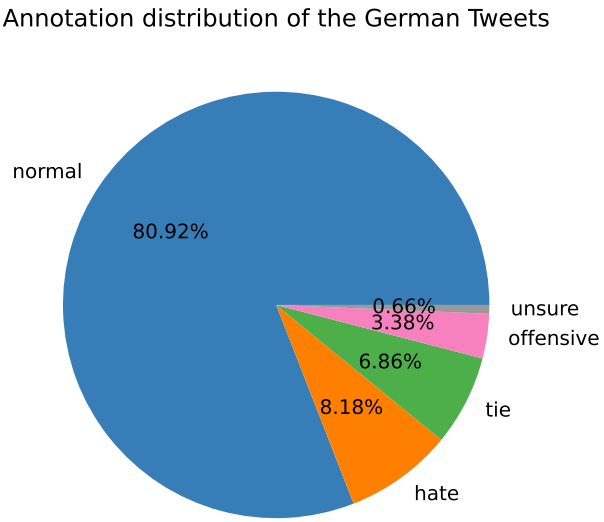

Figure 4.5.: The hate/offensive classification of the German tweets

> @User @User @User und wieder ein Troll zum blocken @PeterMa45118299
> *@User @User and again a troll to block @PeterMa45118299*

This tweet was classified as hate, offensive, normal, both non-racial. The automatic aggregation from Toloka resulted in offensive 84.53% and non-racial target 92.46%. As in this tweet there were a lot of mentions of other users, one can assume that a lot of context is missing which explains the tie. The term *Troll* is an offense but it is unclear to whom the creator writes. So it can be considered a direct offense but also normal speech depending on the context and addressed persons.

Additionally, the single classifications by each annotator were analysed. Comparing two age groups, younger and older than 30 years, the younger group classifies the tweets twice as often as hate (20% compared to 10%) (Figure 4.21). The amounts for unsure and offensive are very similar as well as the classifications for hate and offensive. The main visible difference is that the older annotators are more likely to assign a non-racial target, see Figures 4.22 and 4.23. This already indicates that the age of the annotators should be considered for research.

The annotators who labeled the German tweets have been split by their indicated countries. The country is verified by the phone number. Annotators from Germany tend to classify tweets as hate less often (Figure 4.19) than the other group (4.19% vs 18.16%) (Figure 4.20) and people in Germany are more unsure (5.31% vs 1.83%), see Figure 4.7. They also consider tweets as normal more often (81.26% compared to 70.37%). Another difference is the classification of a racial target. The annotators in Germany consider a third of the hateful tweets as racial whereas the ones not from Germany consider 81.76%.

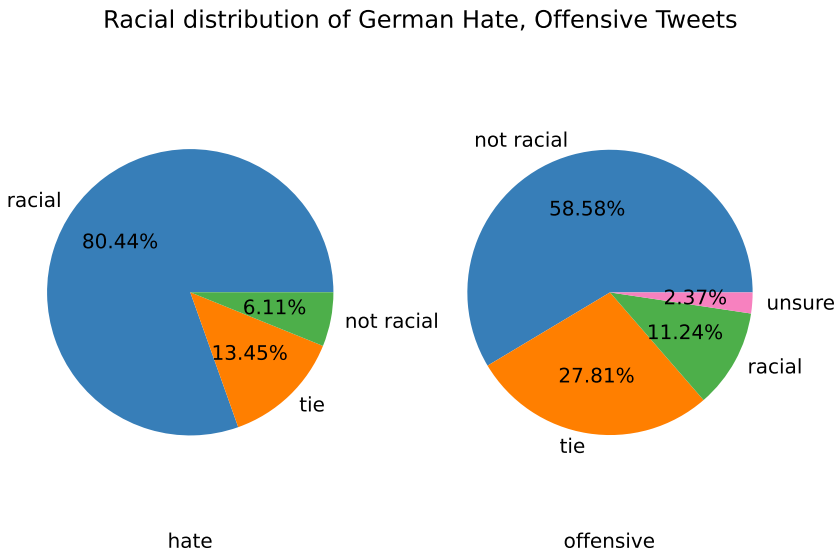

Figure 4.6.: Racial and non-racial targets for both hate and offensive tweets in German

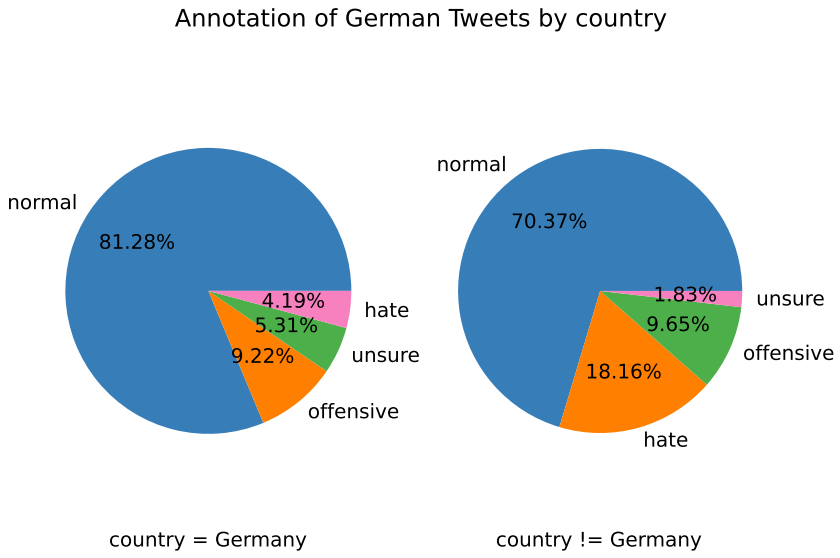

Figure 4.7.: The hate/offensive classification of the annotators depending on the countries

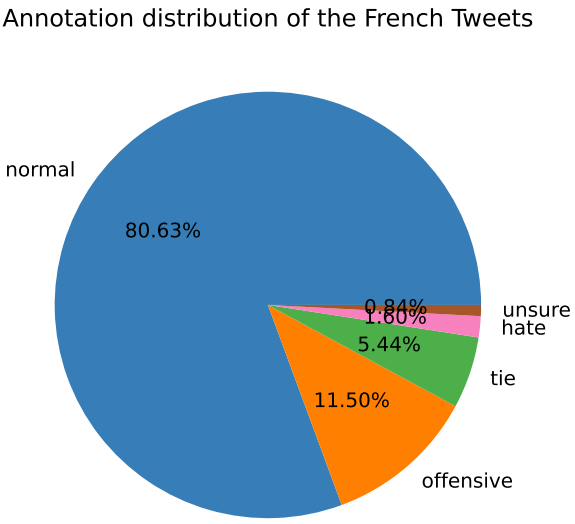

Figure 4.8.: The hate/offensive classification of the French tweets

French tweets

For the French tweets with the majority voting, there is only a very small amount of hateful tweets: 1.6%. This may be explained by the shorter keyword list (see Figure 4.8) which covers a smaller range of topics. 5.44% are a tie, 80.63% are normal and 11.50% are offensive.

The inter-annotator agreement is medium: the Fleiss Kappa[9], which is an agreement score between 0 (poor agreement) and 1 (perfect agreement), of the German dataset is 0.3026, indicating a fair agreement.

Considering the single annotations, there are only very slight differences in the annotations separated by the countries of the annotators as displayed in Figure 4.25 and for the racial and non-racial classification: Figures 4.25 and 4.26. This can be explained as the majority was living in France or Belgium and these countries have less cultural distinctions than the corresponding ones for German or Amharic content. This indicates that the cultural closeness should be considered when choosing the countries to annotate content for other countries/cultural regions.

The same applies to the distinction by age, there are differences smaller than 1% (Figure 4.10). For the racial classification, the older group is more likely to classify the tweet as a non-racial target (by 5%, Figure 4.28 and 4.27). They are also more sure, only around 2% for the hateful and roughly 4% for the offensive tweets are classified as not sure compared to 11.07% and 9.87% respectively for the younger annotators.

Amharic tweets

The Amharic tweets contain 1758 hateful tweets. Due to the lack of available crowdworkers which were not interested in the task when it contained a training pool, this pool could not be used for the Amharic annotation. However, there could still be around 2000 tweets annotated.

[9]Inter-annotator agreement: `https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3`

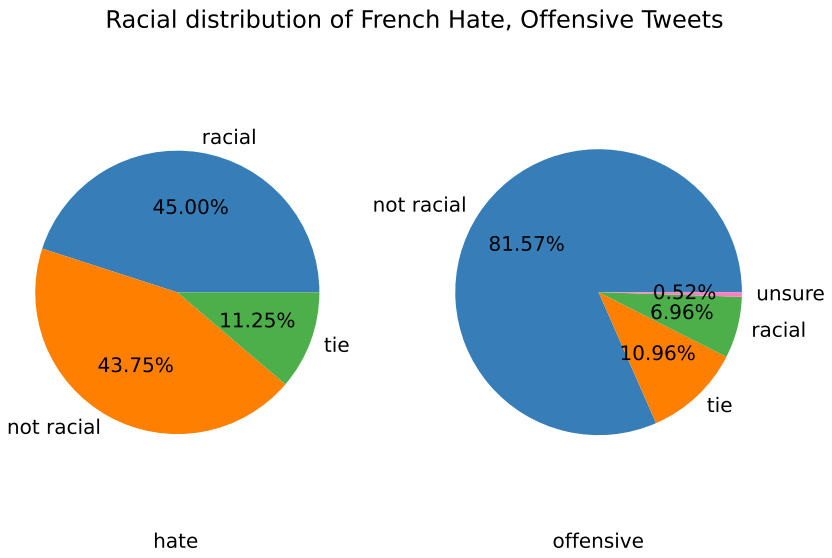

Figure 4.9.: Racial and non-racial targets for both hate and offensive tweets in French

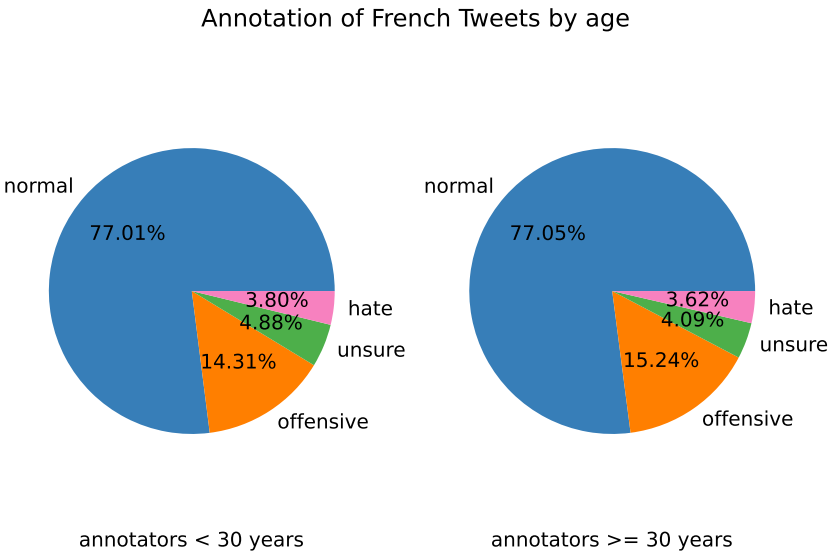

Figure 4.10.: The hate/offensive classification depending on the age

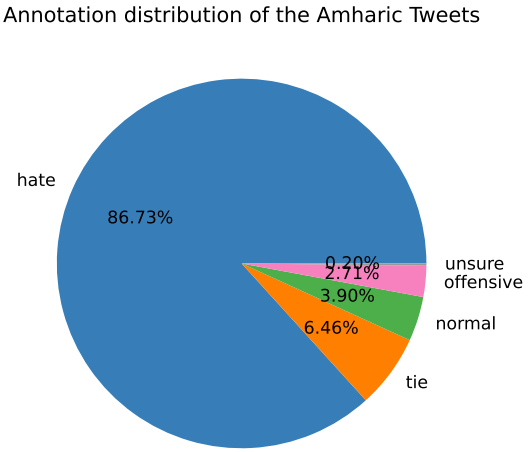

Figure 4.11.: The hate/offensive classification of the Amharic tweets

Among them, there are 3.9 % normal (based on the majority voting), 6.46 % are a tie, 86.73% are hate, 2.71 % offensive and 0.2% unsure, see Figure 4.11. There are fewer ties in the hateful tweets (see Figure 4.12) for the classification of racist tweets and a lot more racial tweets as in the offensive tweets (93.69% vs 36.36%).

The Fleiss Kappa is 0.0918, indicating a low agreement. Due to the high amount of malicious annotators, the quality is not very good.

The classifications considered individually and compared by the age of the annotators are similar. The younger group classifies tweets less likely as hate (24.76% vs 28.71%) and is more unsure (8.83% vs 6.81%), see Figure 4.32. The older group is more sure and found more non-racial targets for the hateful tweets (5.06% not sure, 31.65% nor racial compared to 12.44% not sure and 25.04% not racial, see Figures 4.33 and 4.34). This difference is even larger for the offensive tweets, 16.34% of the younger group is unsure compared to 5.6% and 49.55% were classified as non-racial by the younger group compared to 65.95%.

Comparing the country distributions, there are large differences: People in Ethiopia consider more tweets as normal (50% compared to 31.88%), less as offensive (20.39% compared to 28.16%) and as hate (19.79% compared to 33.44%) and they are more unsure (9.07% vs. 6.52%), see Figure 4.29.

The older annotators group tends to classify tweets less frequent as hate (19.55% compared to 24.76%) and more likely as offensive (28.71% compared to 20.68%), see Figures 4.32. The younger annotators are a bit more unsure (8.83% compared to 6.81%) and the normal classification is almost the same (see Figures 4.33 and 4.34). They are also more unsure considering the racial classification 12.44% for hate and 16.34% for offensive tweets whereas the older ones are unsure for 5.06% and 5.6% of the tweets, respectively. The older group considers less offensive tweets as racial (28.45% vs 34.11%); for the hateful tweets it is they agree more: 63.29% compared to 62.52%.

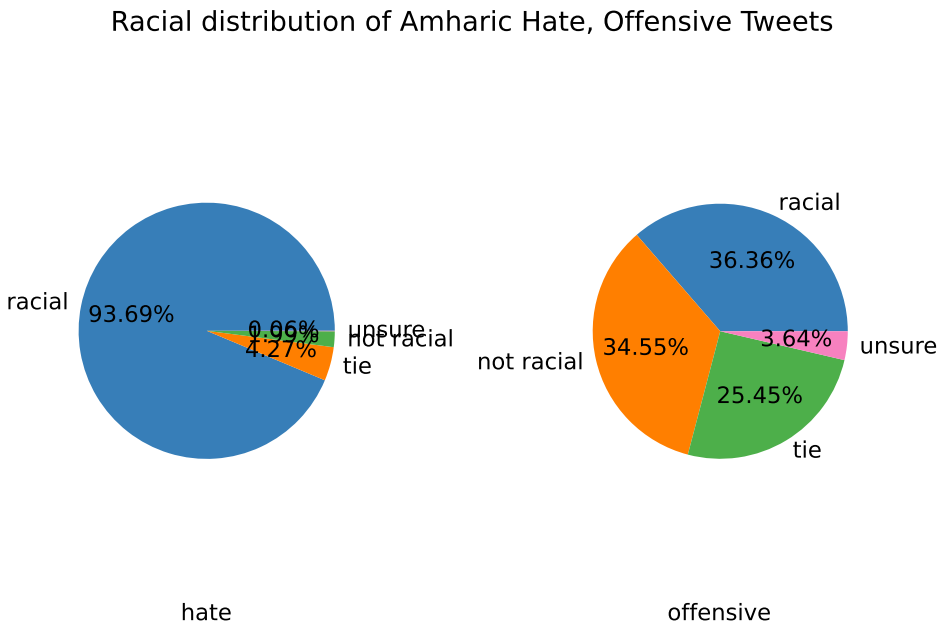

Figure 4.12.: Racial and non-racial targets for both hate and offensive tweets in Amharic

4.2.5. Annotator analysis

According to a survey from Toloka[10] in 2022, there are 245.00 crowdworkers worldwide from 202 countries. The average age is 29.8 years with around 40% female, around 60% male and 0.8% non-binary. The socioeconomic statue varies a lot from the countries but also within them. 83% have graduated from a college. 36% of the Toloka users speak French and 21% German, 98% English.

German-speaking annotators

As there a not enough annotators on Toloka living in Germany or Austria to annotate the data in the time frame, the country restriction was removed. In Figure 4.14, the distribution of the countries is shown. The average age of the annotators is 32.97 years[11], the Figure 4.13 shows the distribution. Some outliers were not considered as Toloka allows impossible ages.

French-speaking annotators

It was sufficient to restrict the countries to France and Belgium as there were enough annotators. However, there were three users from other countries leaving 99% users from France or Belgium (see Figure 4.16). These outliers can be explained by technical issues or by annotators who changed their country after starting the annotation. The average age is 31.11 years, a histogram showing the age distribution can be found in Figure 4.15.

[10]Survey from Toloka: `https://toloka.ai/blog/tolokers-global-survey-2022/`
[11]Github repository, Skadi Dinter: `https://github.com/bickbeermoos/multiling_hatespeech`

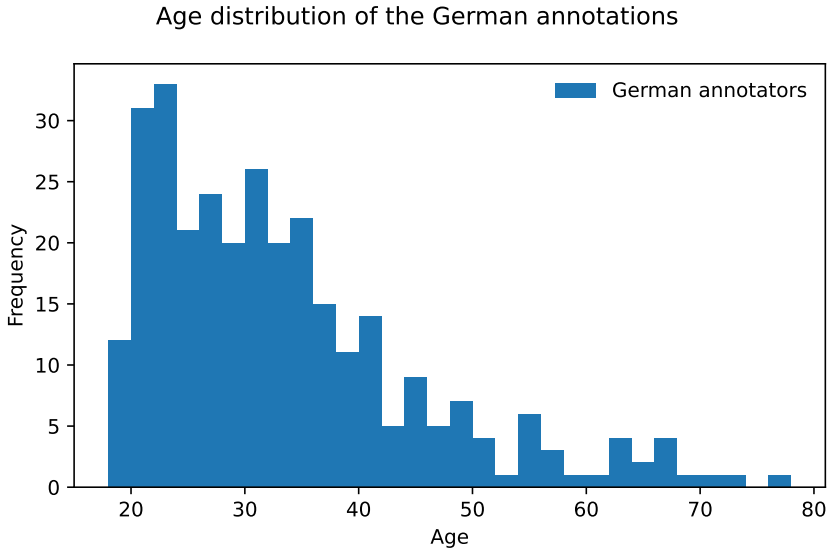

Figure 4.13.: The age distribution of the German annotators

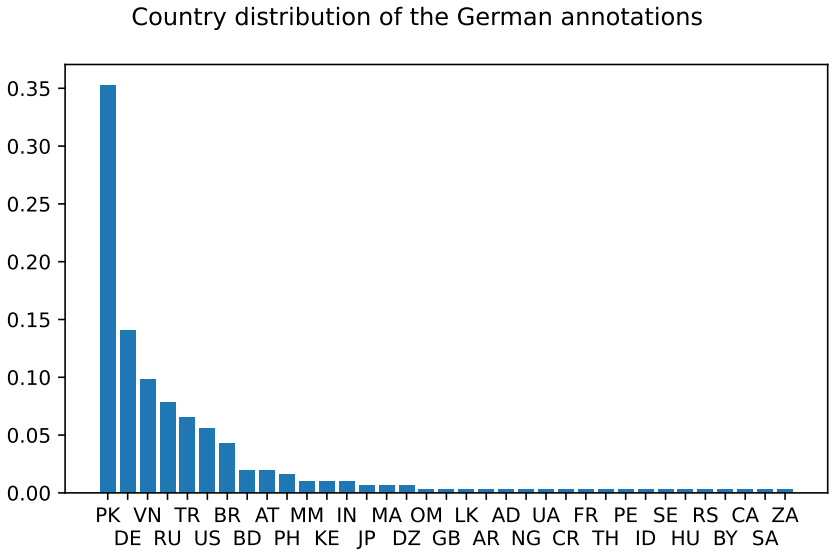

Figure 4.14.: The country distribution of the German annotators

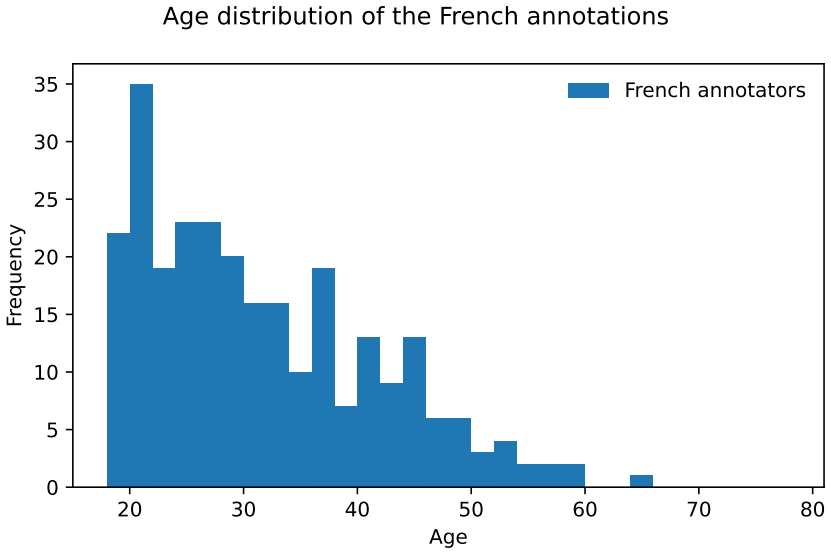

Figure 4.15.: The age distribution of the French annotators

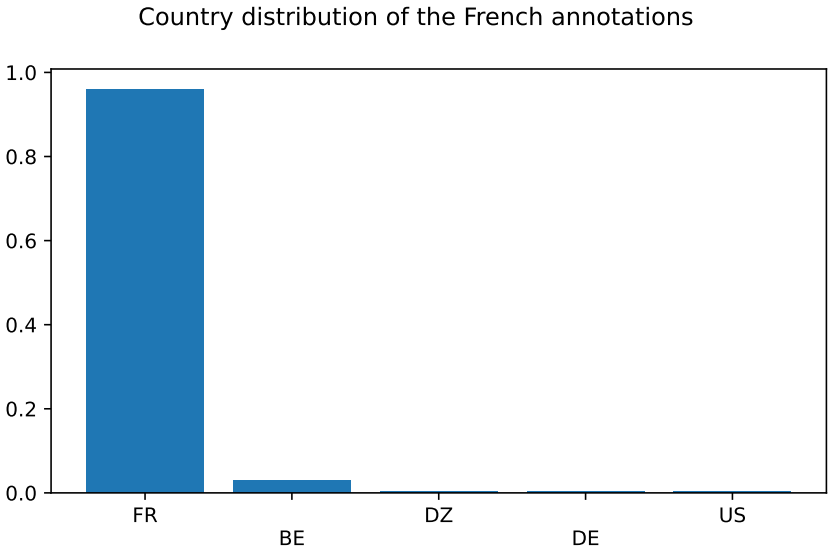

Figure 4.16.: The country distribution of the French annotators

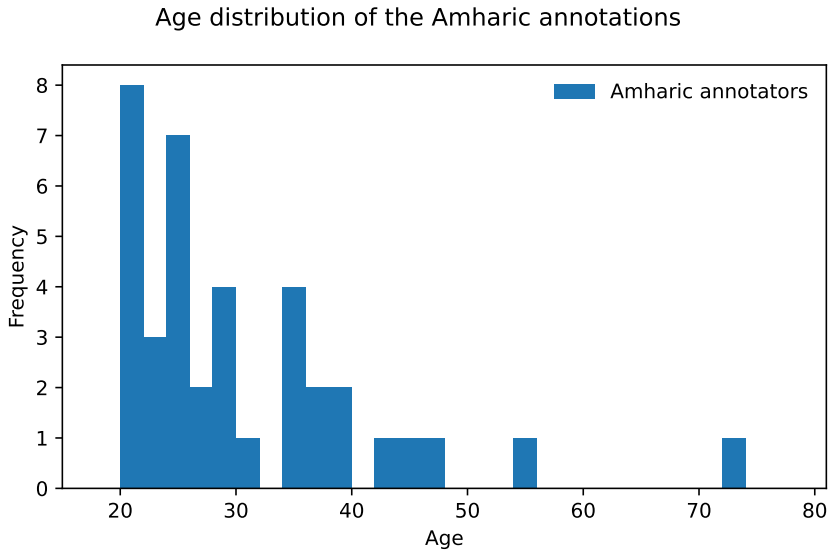

Figure 4.17.: The age distribution of the Amharic annotators

Amharic-speaking annotators

The Amharic annotators are on average 30.24 years old, for the distribution see Figure 4.17. Out of the annotators, 2.6% are from Ethiopia, the distribution is given in Figure 4.18. The majority of the annotators are located in Pakistan. There are not a lot of people who speak Amharic there, indicating that they might not be able to properly understand the meaning of the tweets. The inter-annotator agreement is very low, probably due to the same reasons as for the German annotation.

4.2.6. Figures

In the following, statistics of the annotation and the annotators are presented (Figures 4.19 to 4.34).

4.2.7. Improvements

As the quality of the German and Amharic datasets were not sufficient, possible countermeasures for future studies should be taken into account. As the language tests are very easy to pass, it should be investigated if they can be improved or own language tests included in the training. This would be especially useful for Amharic, as no language test is available on the Toloka platform. As a consequence, it is more clear that the annotators have the required language skills and are able to understand the tweets. Additionally, the size of the tasks in the training pool could be varied to analyse if more example tasks are helpful for the classification. To better understand the classifications of the annotators, there could be an option to indicate the parts of the tweet that were considered important for the classifications, similar to the one in the dataset HateXplain (Mathew et al. 2021). This way, the undecided tweets could be further analysed and handled separately. If more resources are available, it is possible to ask experts with more domain knowledge to classify the tweets. This might lead to higher agreements as they are more

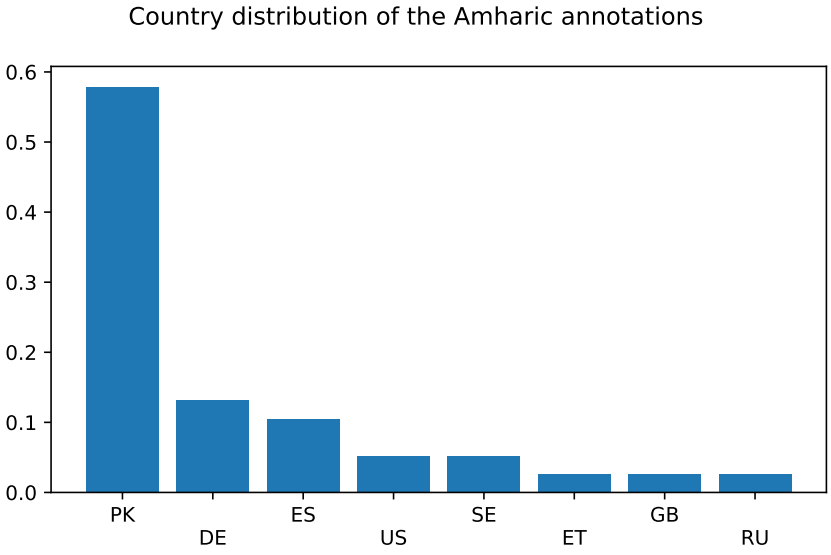

Figure 4.18.: The country distribution of the Amharic annotators

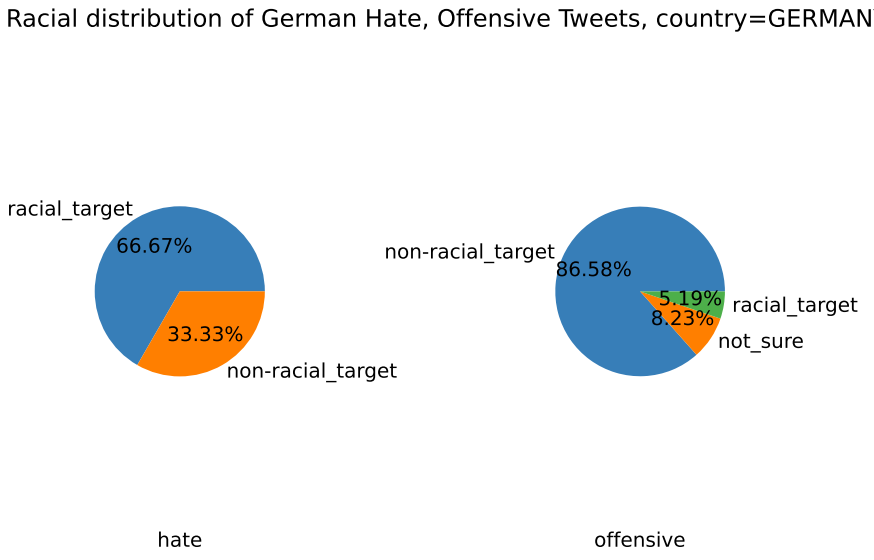

Figure 4.19.: Racial and non-racial targets for both hate and offensive tweets in German where country = Germany

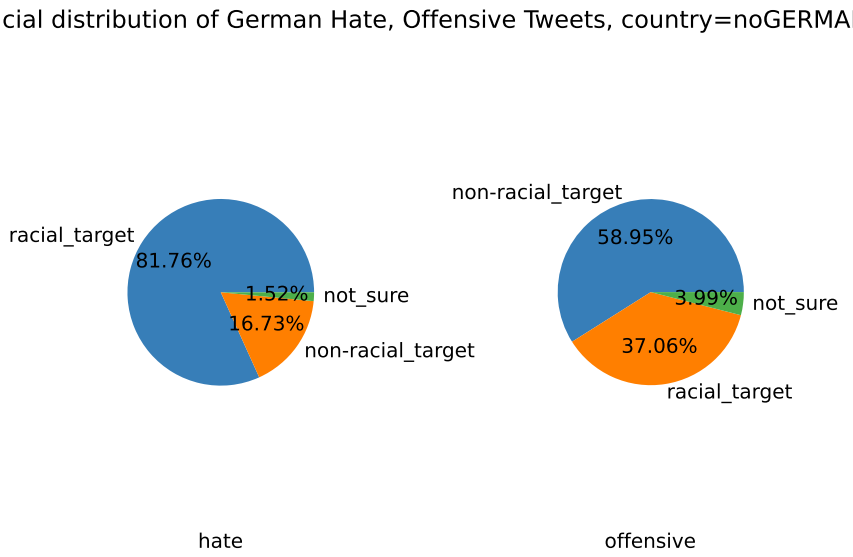

Figure 4.20.: Racial and non-racial targets for both hate and offensive tweets in German where country = noGermany

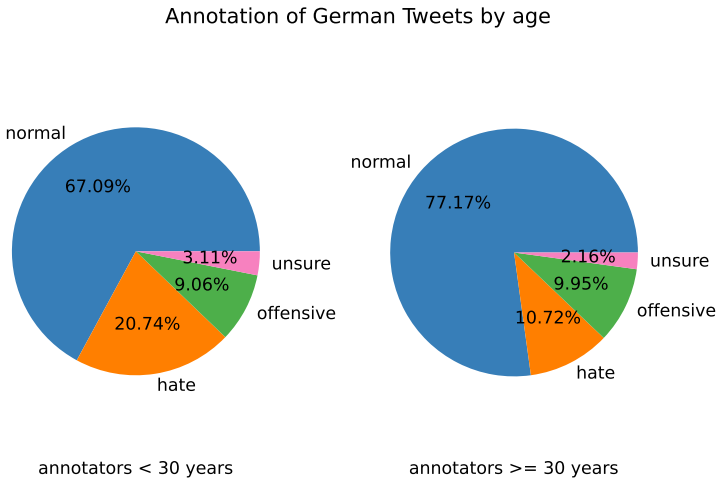

Figure 4.21.: The hate distribution depending on the age

Racial distribution of German Hate, Offensive Tweets of younger annotators

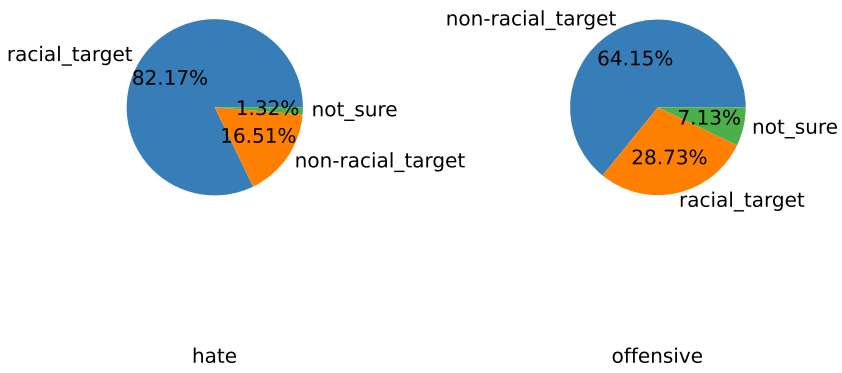

Figure 4.22.: The racial/ non-racial classification of the younger annotators

Racial distribution of German Hate, Offensive Tweets of older annotators

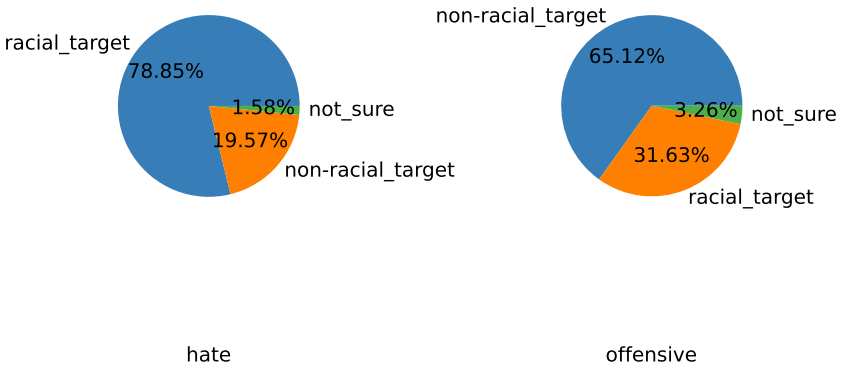

Figure 4.23.: The racial/ non-racial classification of the older annotators

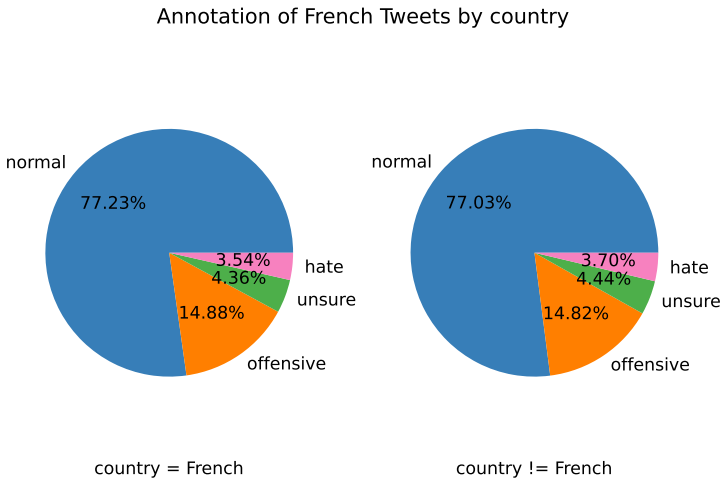

Figure 4.24.: The hate/offensive classification of the French tweets depending on the country

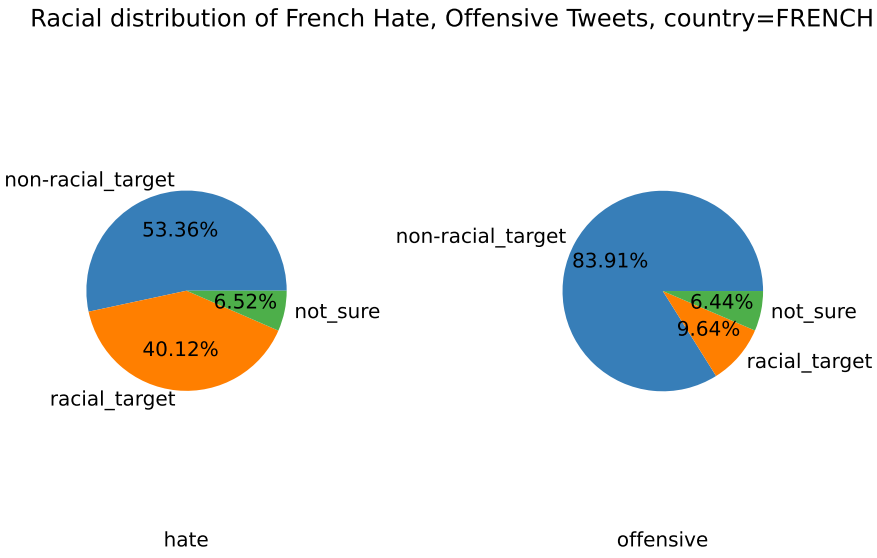

Figure 4.25.: Racial and non-racial targets for both hate and offensive tweets where country= France

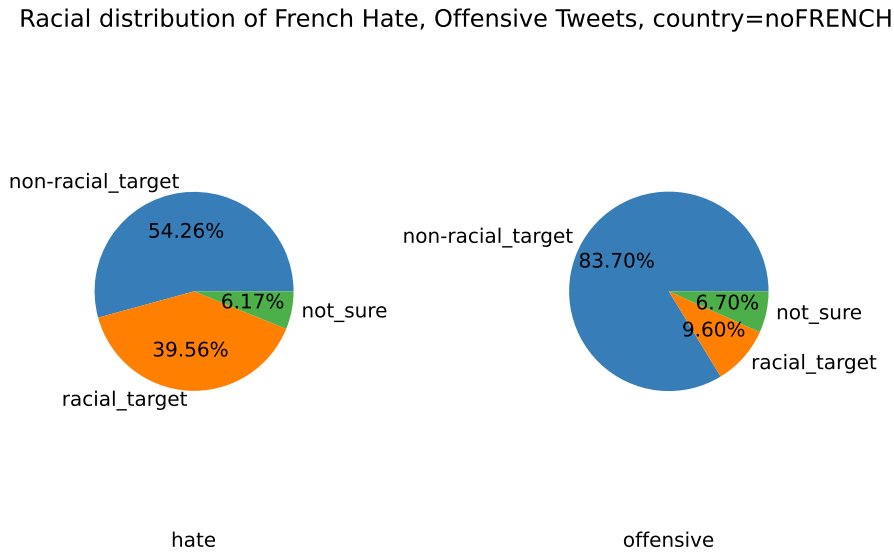

Figure 4.26.: Racial and non-racial targets for both hate and offensive tweets in French where country != France

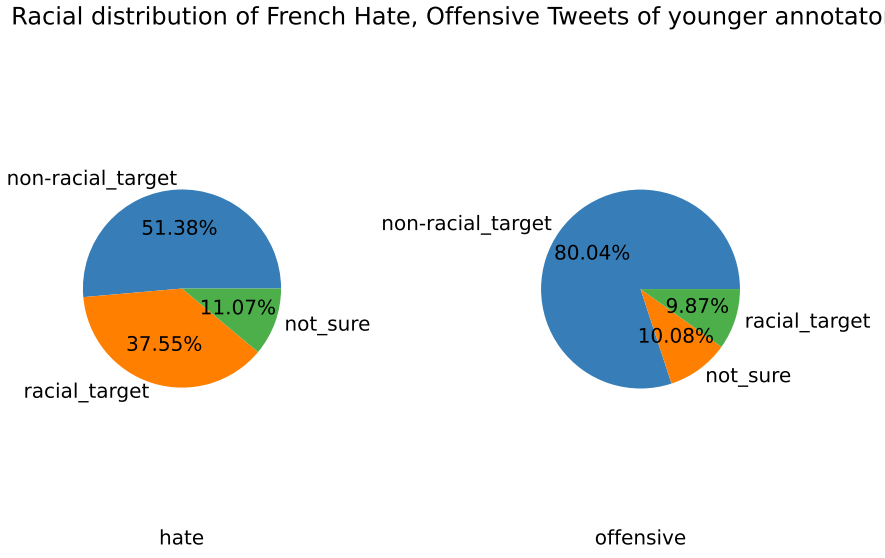

Figure 4.27.: The racial/ non-racial classification of the younger annotators

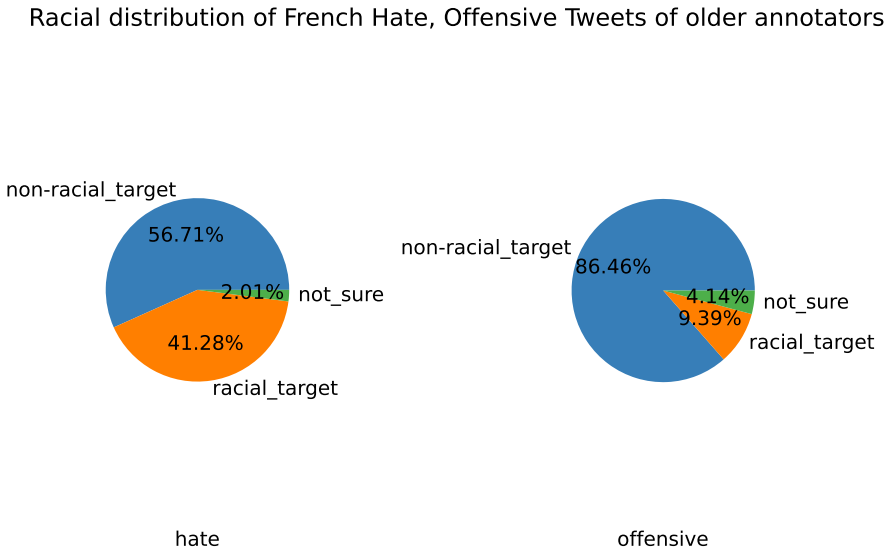

Figure 4.28.: The hate/offensive classification of the older annotators

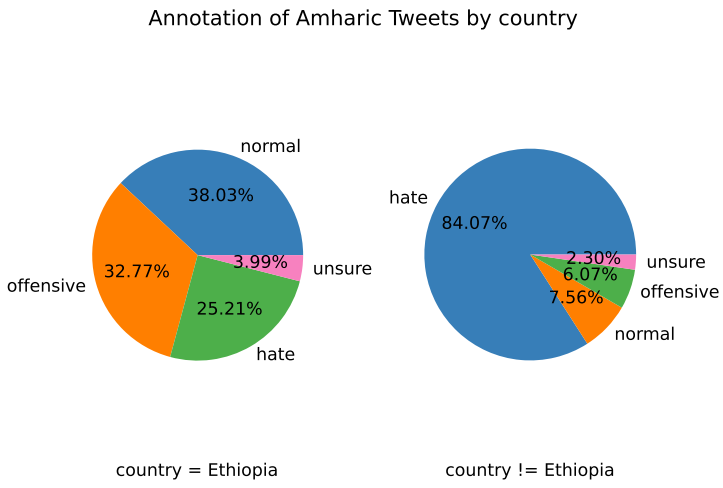

Figure 4.29.: The hate distribution depending on the countries

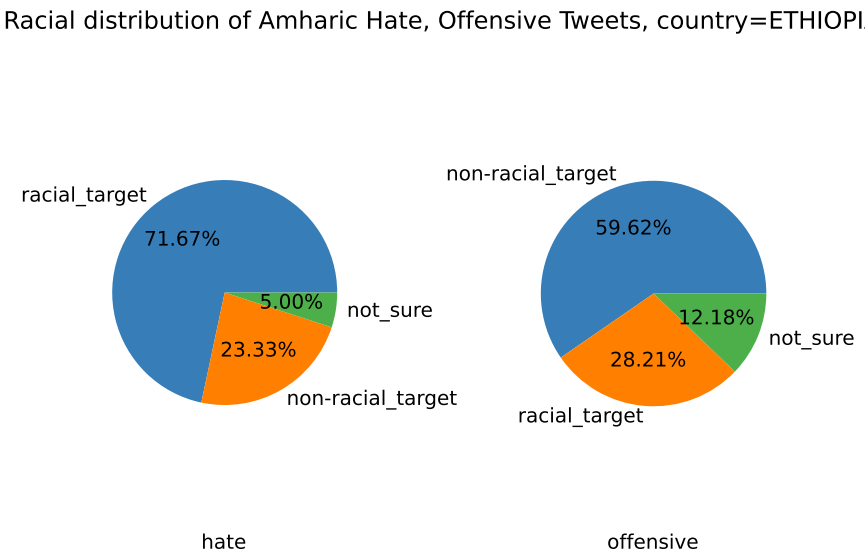

Figure 4.30.: The racial distribution where country = Ethiopia

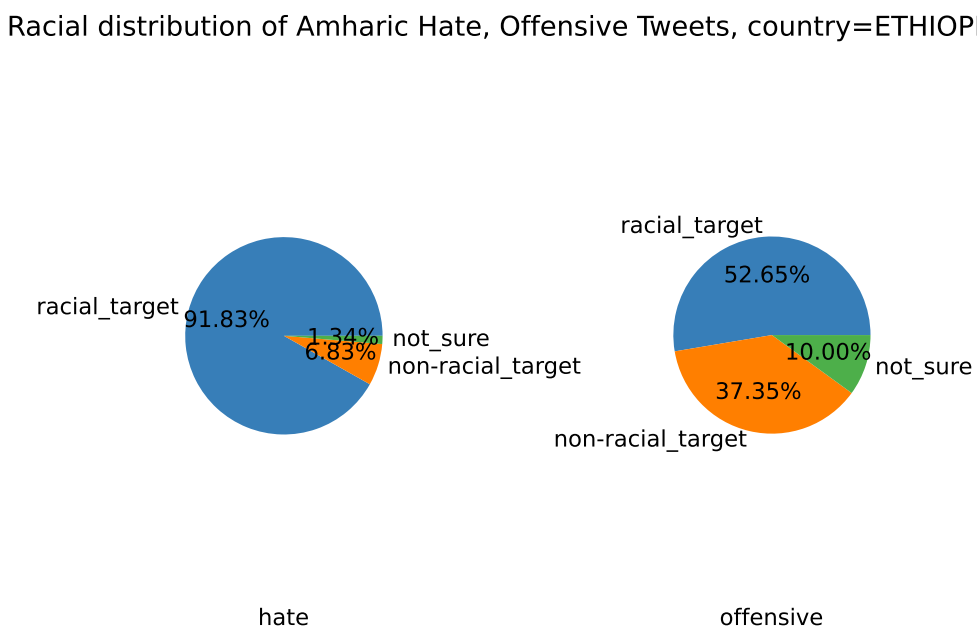

Figure 4.31.: The racial distribution where country !=Ethiopia

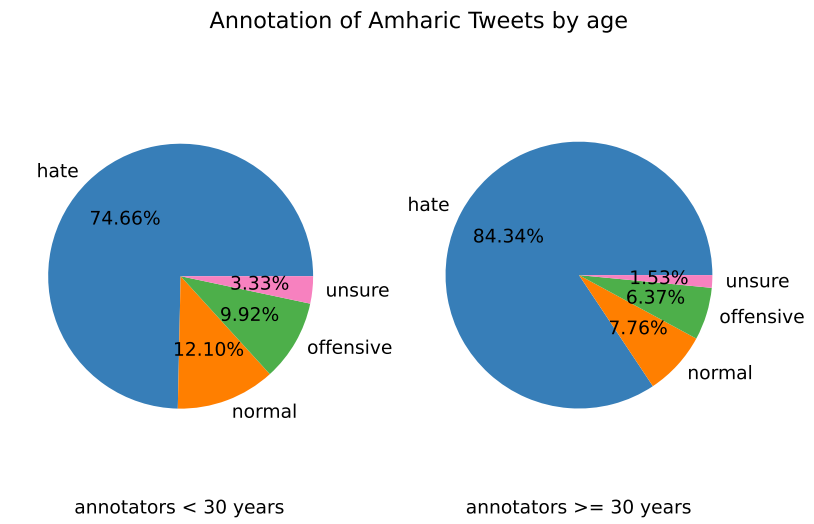

Figure 4.32.: The classification by age

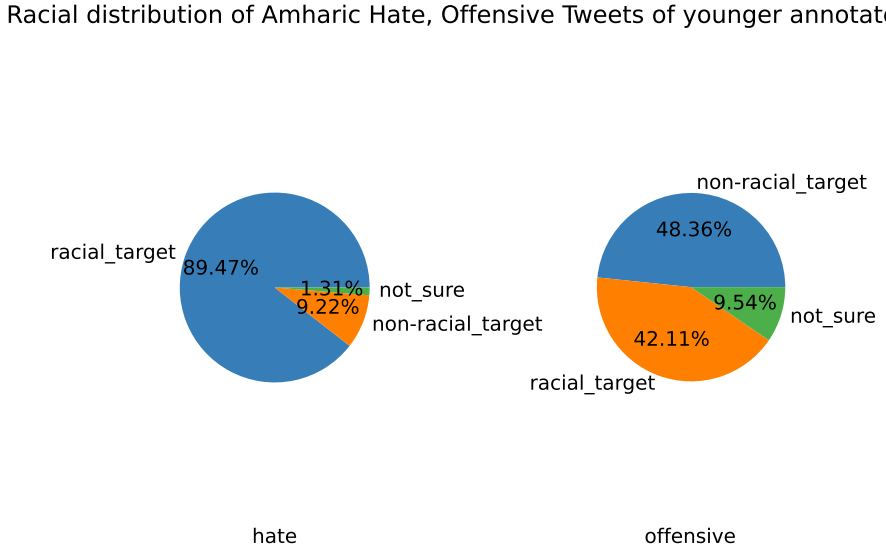

Figure 4.33.: The hate classification of the younger annotators

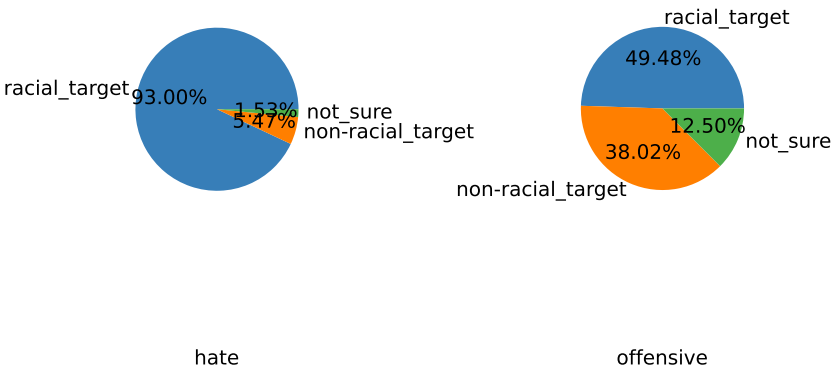

Figure 4.34.: The racial/ non-racial classification of the older annotators

aware of the contexts and slang or codes used in racist language. Another solution would be to investigate if knowing the context of the tweet leads to higher agreement rates, similarly to the work from Gibert et al. (2018), presented in Chapter 3.1. Hence, there are different possibilities for improvements that might lead to higher inter-annotator agreement scores in future studies.

Conclusion

The data collection strategy has been presented as well as the creation of the keyword lists. With the filtering strategy used, around 5000 tweets could be annotated both for French and German and 2000 for Amharic. The statistics show the different agreements and the demographics of the users. A annotation analysis shows a low agreement for German and Amharic; possible reasons are explained. When comparing the languages, one can remark that younger annotators across all languages are more unsure. The French data is used to train a detection model as described in the next chapter.

5. Experimental setup

This chapter introduces the baseline system using HateXplain model and the experiments that were conducted. Some example tweets that were classified by the best model are also analysed.

5.1. Baseline Models

The English language model family BERT includes pretrained models to facilitate natural language processing. They consist of transformer encoder layers with a self attention mechanism (Devlin et al. 2019). The model has grown into a family of language models for a wide range of languages.

HateXplain (Mathew et al. 2021) is a hate speech dataset and detection model based on BERT. The dataset is build from posts on Twitter and Gab which were filtered with keyword lists. It was constructed for data in English and contains rationales to better explain the decisions of the crowdworkers who annotated the posts. The model has achieved an accuracy of 0.698 and a F1-score of 0.687 on this dataset.

5.2. Experiments

For this project, BERT models have been used and fine-tuned with the HateXplain dataset. This allowed to reduce computation costs and not having to train from scratch[1]. Furthermore, compared to creating an own model, the training on the pre-trained model saved time. A smaller dataset is needed and for a variety of tasks good results[2] can be achieved.

The HateXplain dataset was used for finetuning the BERT models which are pretrained for a wide range of language processing tasks. It was further preprocessed and applied for fine-tuning the multilingual BERT model[3]. Additionally, it was translated with Google Translate to French and trained on the French language model camemBERT[4] which is based on the pretrained English transformers model roBERTa[5].

As a next step, the collected and crowd-sourced French dataset was further used to finetune the models. An overview of the various studies conducted with different datasets is given in Table 5.2. For the experiments, the influence of different kinds of datasets was analysed. One of them is an automatic aggregation of the three annotations for each tweet based on the

[1] Fine-tune a pretrained model: https://huggingface.co/docs/transformers/training
[2] BERT finetuning: `https://mccormickml.com/2019/07/22/BERT-fine-tuning/`
[3] BERT multilingual: https://huggingface.co/bert-base-multilingual-uncased
[4] CamemBERT: https://huggingface.co/camembert-base
[5] roBERTa: https://huggingface.co/roberta-base

Dawid-Skene aggregation method[6] as explained in Section 4.2.4 (studies 1.2 and 2.2). Opposed to the automatic aggregation some studies were conducted with a custom aggregation method which combines the votes in the following way: the classifications with at least two votes were considered the ground truth for each tweet. When there are three different classifications, the tweet is either removed (studies 1.1 and 2.1) or if there is at least one hateful label, it is considered hate and otherwise as offensive (studies 1.3 and 2.3).

The models use the following hyperparameters[7] per default:

- Learning rate: 5e-5
- Optimizer: AdamW[8] that uses the Adam algorithm with weight decay regularization (Loshchilov and Hutter 2017)
- Weight decay: 0
- AdamW with beta1: 0.9, beta2: 0.999 epsilon: 1e-8
- Number of training epochs: 3
- Batch size: 8
- Optimization steps: 234
- Training loss: 0.628

5.2.1. Results

For both BERT models used, the datasets perform very similar as it is shown in Table 5.2. Hence, the model based on the Dawid Skene aggregation gained a better accuracy and F1-score than the aggregation based on the ones with a majority voting for both the multilingual BERT and camemBERT. The removal of the votes with ties lead to the best results for both base models. This leads to the conclusion that adding ties does not lead to better results. It would be useful in future research to analyse the ties better and to find more reliable classifications of them to be able to classify them better.

In the first step, the pretrained BERT models which were finetuned with the HateXplain dataset, yielded worse results than the baseline hateXplain model on the same dataset. All studies on the multilingual BERT (1.1) performed worse than the corresponding ones based on camemBERT (2.1). This indicates that it makes more sense to translate English datasets like the HateXplain one and then to use the BERT model that corresponds to the studied language. As soon as the datasets for other languages have acceptable agreement scores, the generalisability should be verified.

[6] Automatic Dawid-Skene aggregation: `https://toloka.ai/docs/guide/concepts/result-aggregation.html`

[7] Hyperparameters of the pretrained BERT models: https://huggingface.co/docs/transformers/main_classes/trainer

[8] AdamW optimizer: https://huggingface.co/docs/transformers/v4.21.3/en/main_classes/optimizer_schedules#transformers.AdamW

Study	Pretrained Model	Label generation	Accuracy	F1-score	Ties	Training time	Samples
1.0	ML BERT	HateXplain	0.51	0.41	-	12m 47s	20149
1.1	ML BERT+ HateXplain	self aggregated	0.84	0.77	no ties	3m6s	4728
1.2	ML BERT+ HateXplain	Dawid Skene	0.78	0.69	automatically	4m3s	5000
1.3	ML BERT+ HateXplain	self aggregated	0.65	0.51	if hate: hate, otherwise offensive	4m9s	5000
2.0	camemBERT	HateXplain	0.592	0.57	-	10m45s	20149
2.1	**HateXplain on camemBERT**	**self aggregated**	**0.888**	**0.86**	**no ties**	**3m19s**	**4728**
2.2	HateXplain on camemBERT	Dawid Skene	0.806	0.75	automatically	3m54s	5000
2.3	HateXplain on camemBERT	self aggregated	0.726	0.674	if 1 hate: hate, otherwise offensive	3m12s	5000

Table 5.1.: Studies for building a French hate speech detection model based on different BERT models and datasets

5.2.2. Predictions

The Table 5.2 details example classifications. The offensive tweets were classified well but some normal tweets were also classified as offensive. The racial labels were not considered yet and will need to be taken into account for training in future studies.

There are remarkable differences between the performance of the models based on the multilingual BERT and the French camemBERT. Whilst the multilingual BERT always predicts *normal* as the class label with nearly the same score for every tweet, the camemBERT labels the tweets appropriately. The multilingual studies achieve a lower score than the camemBERT models, repsectively. There remain misclassifications as shown in Table 5.2. The size of the classes for a random sample of 50 misclassifications is given together with an analysis on why this tweet might have been misclassified. Even for tweets where all three annotators agreed, there are misclassifications. For example, no tweet in the test set was classified as hate even though there were examples from annotators who all agreed that the corresponding tweet was hateful. This can be explained by the class imbalances in the original dataset as only 1.6% of the tweets were classified as hateful.

The two main differences in the base model, namely the camemBERT and the Multilingual BERT model lead to different results. The models predict normal for the test set which is not expected and can be explained by the class imbalances in the dataset. As described in Section 4.2.4, the amount of hateful tweets in the French dataset is 1.6% compared to 80% normal speech.

As no examples were classified as hate, future research is needed to collect more samples to train on. Alternatively, one could experiment with different sampling weights for the hate class.

5.2.3. Further finetuning

For both the multilingual BERT and camemBERT, the best performing finetuned model was chosen and hyperparameters like the number of epochs and the learning rate were varied as shown in Table 5.3 and 5.4. As the dataset has unbalanced classes, a stratified splitting of both the train and the test set was chosen as another experiment.

For the model based on the multilingual BERT, there were no noticeable differences in the experiments. The accuracy and the F1-score did not change and the predictions were still always the same. For the camemBERT-based model however, the accuracy could be improved by 3 percent and the F1-score by 0.01. The predictions did not change but the confidence score increased from 0.01 to almost 0.2.

Conclusion

The baseline model HateXplain was presented. Different methods to classify French hate speech were presented and evaluated. It was shown that is possible to fine-tune an existing BERT model with the translated HateXplain dataset. Predictions of the best working model were shown to work well on normal speech. However, due to the class imbalances the predictions are not yet accurate for all cases. Some misclassifications were further analysed.

Error class	Size	Tweet	Label (score)	Ground truth	Analysis
FN Offensive	29/50	@User T'a pris des cours de français avec jawed comme prof ou quoi ? *@User Did you take French classes with jawed as a teacher or what?*	Normal (0.652)	Offensive	The offense against a person is hidden making it more difficult to classify
FN Hate	6/50	@User Et celui qui filme lái aussi cést putain de blacklivesmatter le sont tous. *@User And whoever is filming it has it too it's fucking blacklivesmatter are all.*	Normal (0.739)	Hate (all three)	The hate is expressed directly, but context information is needed to understand it
FN Hate	6/50	Au lieu de s'occuper de Marine Le Pen le maire de Dijon ferait mieux de s'occuper des dealers maghrébins et des gangsters tchetchens *Instead of dealing with Marine Le Pen the mayor of Dijon would do better to deal with North African drug dealers and Chechen gangsters*	Offensive (0.448)	Hate	The problematic speech was recognised but not hate
FN Normal	8/50	@User Bah faut avoir un mec *@User Bah must have a guy*	Offensive (0.9239)	Normal	There is no offense but the use of informal language
Ties	6/50	@User @User Pour le coup ses congénères ont l'air moins bête que lui *@User @User At the moment, his fellows look less stupid than him*	Normal (0.818)	Tie	It was a tie for the annotators: normal, unsure and offensive were selected (a tie is not aimed for in the classification model)
Unsure	1/50	@User @User C'est surtout de toi que les gens devrait se désabonner murji que tu es *@User @User It's mostly you that people should unsubscribe from murji that you are*	Normal (0.923)	Unsure (1 offensive, 2 unsure)	More context information is necessary to decide if the tweet was classified properly
FP Normal	33/50	TW scène de violence d'agents sncf sur une personne noire. *TW scene of violence of agents sncf on a black person*	Normal (0.887)	Hate	Here the hate is indirect, more context is necessary
FP Offensive	17/50	il a dû invoquer la voix d'une intelligence artificielle pcq personne voudrait réciter un texte aussi absurde *he had to invoke the voice of an artificial intelligence because nobody would recite such an absurd text*	Offensive (0.904)	Normal	Not clear why this was misclassified

Table 5.2.: Analysis of example predictions with the score and (size of) error class by the model from study 2.1

Study	Accuracy	F1	Time to train	Samples	Epochs	Learning rate
1.1 a)	0.852	0.784	3m30.847s	4728	3	5e-5
1.1 b)	0.852	0.784	1m13.524s	4728	1	5e-5
1.1 c)	0.852	0.784	2m24.933s	4728	2	5e-5
1.1 d)	0.852	0.784	2m24.933s	4728	3	5e-4
1.1 e)	0.852	0.784	3m51.095s	4728	3	5e-6

Table 5.3.: Experiments based on the study 1.1 with varied epochs and learning rates, all use stratified splitting

Study	Accuracy	F1	Time to train	Samples	Epochs	Learning rate
2.1 a)	0.886	0.859	4m53s	4728	3	5e-5
2.1 b)	0.899	0.882	2m12s	4728	2	5e-5
2.1 c)	0.888	0.876	1m14s	4728	1	5e-5
2.1 d)	0.882	0.869	4m10s	4728	4	5e-5
2.1 e)	0.852	0.784	3m11s	4728	3	5e-4
2.1 f)	**0.892**	**0.869**	**3m13s**	**4728**	**3**	**5e-6**
2.1 g)	**0.892**	**0.874**	**3m56s**	**4728**	**4**	**5e-6**

Table 5.4.: Experiments based on the study 2.1 with varied epochs and learning rates, all use stratified splitting

6. Discussion

In this chapter, the results of the dataset collection and model building are presented. The cultural context is considered for the French dataset and patterns of racist language are compared to the tweets. Limitations of my work and possibilities for future work are described.

6.1. Results

For this thesis, three datasets could be collected: a German, a French and an Amharic one. They were annotated by crowdworkers with the possible categories: hate, offensive, normal or unsure. For the first two options, the annotators could choose if the tweet is directed against a racial target or not or if they are unsure. However, only the French dataset had a sufficient inter-annotator agreement score. Possible explanations for the low results in German and Amharic are given in Section 4.2.4. The other datasets and their annotator groups were also analysed.

The French dataset contains the least amount of hate compared to the other languages. This may be because of the low amount of keywords covering a smaller variety of hateful words or because in their culture it is rare to speak about racism or hate (Wang et al. 2021). However, the agreement is very high in French. Thus, a lot of annotators agreed that a tweet is normal which is the easiest category to identify and the one with naturally the most occurrences (as the overall amount of problematic content is low).

The amount of keywords seems to not correlate with the quality of the dataset. The German keyword lists contains 17367 entries, the Amharic one 147. Both achieved low agreement rates and thus were not used to build any models. The French list however contains 3473 keywords.

As described in Section 4.2.4, the age of the annotators correlates with different classifications for the German tweets. Thus, for better results it might be useful to find annotators with a similar age range than the user demographics of the platform that is investigated. Using the results from the German annotation, it might make sense to better describe racial targets and how to classify them as the two age groups classified the tweets differently in this dimension.

As the agreements were to low, only a French hate speech detection model based on BERT models and the HateXplain dataset was build as described in section 5.2. Compared to the HateXplain model, the performance metrics are worse but the ones specific for this task are higher. Normal and offensive speech can be predicted most often correctly for example tweets. There remain some challenging classifications however, in particular regarding hate speech, as shown in Table 5.2. This can be explained by class imbalances in the dataset.

6.1.1. Language and cultural analysis

Racial language typically covers only a limited amount of topics like migration, crime and economy. Racist discourse also consists of stereotypes, prejudiced statements and truth claims. Specific n-grams are also used more often, for example *our own kind* which usually refers to a specific group in distinction to a group that is discriminated against. Hence, while expressing racist hate, there is usually one group referred to as "us", which is considered normal and opposed to another group, called "them" (Tulkens et al. 2016). This other group is usually used for another ethnic group but it can refer to all kinds of discrimination as described in the definition in Section 2.1.

Some of these characteristics could also be found in the tweets that were classified as racial:

> @User cki tous ces pd de maghrébin faible qui le suis
> *@User cki all these weak North African pd who follow him*

This tweet was classified as racial and offensive by a majority voting. One annotator chose hate instead of offensive and all 3 chose the racial target. Thus, they were very sure, probably because North Africans are directly offended in this tweet.

Exactly the same classifications were chosen for the following tweet:

> Maintenant la guerre raciale est évidente :) (Arabe Vs Négre) Vs Blanc :) #PopCorn le flic avait raison @A
> *Now the race war is obvious :) (Arab And Negro) And White :) #PopCorn the cop was right A*

This can be explained by the direct use of race words, violence (*war*) and the comparison of the groups, indicating a distinction. The second sentence however is not understandable without its context.

One example for racial hate speech is:

> ça devient du grand n'importe quoi. bientôt va falloir qu'on s'excuse d'être né blanc.
> *it becomes big anything. soon we will have to apologize for being born white.*

This tweet includes a truth claim and a construction of us and them.

As described in Section 2.3.2, racism is a rarely discussed topic but widespread problem in France. The killing of George Floyd and the following protests as well as the beginning of the Covid19 pandemic have accelerated racist behavior and protests against racism at the same time. Regarding the annotation, the workers were very hesitant and annotated 1.6% of the tweets as hateful. This might be explained by the lacking awareness of hate and racism and public debates as well as by the filtering of the tweets which did not include enough hateful tweets. Among those tweets classified as hateful, none were considered unsure but 11.25% were a tie. This further indicates that the classification is a difficult task and that it can be very subjective. For the offensive tweets, 0.52% were unsure and 10.96% a tie. These ties should be analysed extensively in future research as they can give insights into challenging tweets. The second explanation will be analysed in more detail in the following section. There are only five tweets

mentioning George Floyd and four were classified as normal and one as undecided. A possible explanation is that police violence and racism are not considered a big problem in France in public debates and thus it was not as often discussed in French tweets. Around 5% of the annotated tweets were undecided indicating that there are still some tweets which are difficult to classify or that they do not agree on common definitions. The French annotation was similar among the age groups. These results should be tested with another dataset containing more hate to confirm these findings.

As described in Section 4.2.4, the similarity between annotators from France and Belgium led to very similar classifications in French. The German and Amharic tweets were annotated by workers from some countries that have a very low amount of people that speak German or Amharic. This indicates that the classifications differ a lot by country. Future research is needed to investigate the actual language skills of the annotators to confirm that the linguistic region influences the annotation results.

6.1.2. Research question

The research question presented in Chapter 1.1 was:

> **Can we create a hate speech detection model (based on BERT and HateXplain) that can be efficiently adapted to other languages or cultures, specifically German, French and Amharic?**

As discussed in this chapter, it was possible to collect and annotate data for hate speech detection. It was not possible to get a sufficient dataset quality for German and Amharic. For the French dataset however, several hate speech detection models that were based on different BERT models and the HateXplain dataset could be created. By using the fine-tuning method, it was possible to increase the confidence scores of the classifitation and the gained accuracy was higher than the HateXplain model. However, there remain several biases and limitations. The main emphasis was to detect racial hate speech and offenses with this model.

6.2. Limitations and future research

Twitter is very often studied, in the study from Vidgen and Derczynski (2020) it was used in more than the majority of the investigated studies (55%), probably due to the easily accessible API (Mathew et al. 2021). Twitter may thus be over-used (Vidgen and Derczynski 2020) making the language model less generalizable to content from other social media platforms. Furthermore, as a tweet contains at most 280 characters, the generalisabilty is limited for longer texts. Even though retweets were not considered, there remain tweets that need context or background information such that their meaning can be understood and they can be properly classified.

The Twitter users' demographics are not representative of the whole population. In general, Twitter users are wealthier and younger than people that do not use it (Vidgen and Derczynski 2020). These differences may bias the collected data (Vidgen and Derczynski 2020).

Furthermore, the class distribution needs to be considered when evaluating a model's output (Vidgen and Derczynski 2020). It is difficult to find a lot of hate on Twitter as in most online communities there is less than 1% of abusive comments. It was not possible to collect a lot of abusive data in the French dataset, one way to improve the class balance would be to investigate

the lexicon methods further. The amount of keywords varied greatly over the the different languages but it could give no indication on the influence of the quantity to the agreement. Future research is needed to investigate which of the French keywords were in particular helpful to find hateful tweets. The role of keywords should be considered as they can introduce new biases and hate can be expressed without keywords. In order to find other structures of hate that ease the filtering of tweets for the data collection, more detailed linguistic research could be helpful.

Regarding the annotation process, there remain several limitations. Currently, there is no way to validate that the annotators read and understood the guideline. Even though the guidelines were used to educate and guide the annotators, they may not have been read or were too long or unspecific. Due to the resources and time constraints, the annotators were not necessarily experts which might have influenced the quality. The task itself, distinguishing between hate and offensive content, is very difficult as in all languages the annotators were sometimes unsure or there were ties as they disagreed on the classification.

To investigate the quality of the hate speech detection model further, the datasets in Section 3.1 can be used for further comparisons between the data sets.

Other dimensions of discrimination and intersectionality should be analysed in future research. In this thesis, hate speech and racist debates were discussed for the United States but no dataset has been collected. Thus, it would be useful to study extensively the existing datasets or collect a new one to compare English tweets to the German, French and Amharic ones.

Conclusion

In this chapter, the overall was summarized and the important results were presented. They were placed into context of the linguistics of racial hate speech and the culture in France. The research question was answered and limitations and possibilities for future research discussed.

7. Conclusion

Hate speech as a challenging problem has been introduced and defined. Different kinds of hate have been presented as well as offensive speech and harassment. It is difficult even for humans to distinguish these problematic speeches as hate can be expressed in a very subtle way, hidden in codes or depend on the context. Racism is one of the most common dimensions of hate. The killing of George Floyd on May 25th, 2020, has started debates on racism and racial profiling but also accelerated racist utterings on social media. Motivated by the severe consequences, both for the society and for individuals, different ways to combat this have been developed. Automatic detection algorithms have been developed and presented with their current state and limitations. Most of the existing classifiers only analyse English content. As most of the content in social media is not in English, there is a need to recognise hate also in other languages.

A research question was hence derived:

> **Can we create a hate speech detection model (based on BERT and HateXplain) that can be efficiently adapted to other languages or cultures, specifically German, French and Amharic?**

Related works have been presented for data collection and annotation methods and content in the various languages. A dataset collection of hate speech examples is given and possible biases as well as few comparative studies were introduced.

The dataset collection for this thesis has been introduced, including the annotation process with crowd-sourcing. The results from the annotation were discussed. As the German and Amharic dataset did not yield sufficient inter-annotator agreement scores, only the French dataset (kappa of 0.3) was further used to fine-tune a BERT model together with the translated HateXplain dataset. The best study gained an accuracy of 0.88 which is better than the baseline HateXplain model. The results from the annotation and the model creation were discussed and limitations of this work analysed. Possibilities for future work were also given: the annotators' language skills should be verified, the annotators demographics should be compared to the ones from Twitter. Other discrimination dimensions and intersectionality should also be taken into account.

8. Bibliography

Alkomah, Fatimah and Xiaogang Ma (2022). 'A Literature Review of Textual Hate Speech Detection Methods and Datasets'. In: *Information* 13.6. ISSN: 2078-2489. DOI: 10.3390/info13060273. URL: https://www.mdpi.com/2078-2489/13/6/273.

Aluru, Sai Saketh, Binny Mathew, Punyajoy Saha and Animesh Mukherjee (2020). 'Deep Learning Models for Multilingual Hate Speech Detection'. In: *CoRR* abs/2004.06465. arXiv: 2004.06465. URL: https://arxiv.org/abs/2004.06465.

Bai, Xiaoyu, Flavio Merenda, Claudia Zaghi, Tomasso Caselli and Malvina Nissim (2018). 'RuG at GermEval: Detecting Offensive Speech in German Social Media'. English. In: *Proceedings of the GermEval 2018 Workshop*. Ed. by Josef Ruppenhofer, Melanie Siegel and Michael Wiegand. 14th Conference on Natural Language Processing. KONVENS 2018 ; Conference date: 19-09-2018 Through 21-09-2018. Austrian Academy of Sciences, pp. 63–70.

Basile, Valerio, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso and Manuela Sanguinetti (June 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*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 54–63. DOI: 10.18653/v1/S19-2007. URL: https://aclanthology.org/S19-2007.

Beaman, Jean and Jennifer Fredette (2022). 'The U.S./France Contrast Frame and Black Lives Matter in France'. In: *Perspectives on Politics*, pp. 1–16. DOI: 10.1017/S1537592722001104.

Boutwell, Brian B., Joseph L. Nedelec, Bo Winegard, Todd Shackelford, Kevin M. Beaver, Michael Vaughn, J. C. Barnes and John P. Wright (2017). 'The prevalence of discrimination across racial groups in contemporary America: Results from a nationally representative sample of adults.' In: *PloS one* e0183356. URL: https://doi.org/10.1371/journal.pone.0183356.

Breitfeller, Luke, Emily Ahn, David Jurgens and Yulia Tsvetkov (Nov. 2019). 'Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts'. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 1664–1674. DOI: 10.18653/v1/D19-1176. URL: https://aclanthology.org/D19-1176.

Bretschneider Uwe; Peters, Ralf (2017). 'Detecting Offensive Statements towards Foreigners in Social Media'. In: *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS)*. URL: https://pdfs.semanticscholar.org/23dc/df7c7e82807445afd9f19%5C%5C474fc0a3d8169fe.pdf.

Chiril, Patricia, Farah Benamara, Véronique Moriceau, Marlène Coulomb-Gully and Abhishek Kumar (July 2019). 'Multilingual and Multitarget Hate Speech Detection in Tweets'. In: *Conférence sur le Traitement Automatique des Langues Naturelles (TALN - PFIA 2019)*. Toulouse, France: ATALA, pp. 351–360. URL: https://hal.archives-ouvertes.fr/hal-02567777.

Chiril, Patricia, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi and Marlène Coulomb-Gully (2020). 'An Annotated Corpus for Sexism Detection in French Tweets'. In: *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1397–1403.

Chung, Yi-Ling, Elizaveta Kuzmenko, Serra Sinem Tekiroglu and Marco Guerini (July 2019). 'CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 2819–2829. DOI: 10.18653/v1/P19-1271. URL: https://aclanthology.org/P19-1271.

Corazza, Michele, S. Menini, Elena Cabrio, Sara Tonelli and S. Villata (2020). 'A Multilingual Evaluation for Online Hate Speech Detection'. In: *ACM Transactions on Internet Technology (TOIT)* 20, pp. 1–22.

Davidson, Thomas, Debasmita Bhattacharya and Ingmar Weber (2019). 'Racial Bias in Hate Speech and Abusive Language Detection Datasets'. In: *CoRR* abs/1905.12516. arXiv: 1905.12516. URL: http://arxiv.org/abs/1905.12516.

Davidson, Thomas, Dana Warmsley, Michael Macy and Ingmar Weber (2017). *Automated Hate Speech Detection and the Problem of Offensive Language*. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15665/14843.

Demilie, Wubetu Barud and Ayodeji Olalekan Salau (2022). 'Detection of fake news and hate speech for Ethiopian languages: a systematic review of the approaches'. In: DOI: https://doi.org/10.1186/s40537-022-00619-x.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2019). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain and Lucy Vasserman (Dec. 2018). 'Measuring and Mitigating Unintended Bias in Text Classification'. In: pp. 67–73. DOI: 10.1145/3278721.3278729.

ElSherief, Mai, Vivek Kulkarni, Dana Nguyen, William Yang Wang and Elizabeth M. Belding (2018). 'Hate Lingo: A Target-based Linguistic Analysis of Hate Speech in Social Media'. In: *CoRR* abs/1804.04257. arXiv: 1804.04257. URL: http://arxiv.org/abs/1804.04257.

Fangen, Katrine and Lisanne Lichtenberg (2021). 'Gender and family rhetoric on the German far right'. In: *Patterns of Prejudice* 55.1, pp. 71–93. DOI: 10.1080/0031322X.2021.1898815. eprint: https://doi.org/10.1080/0031322X.2021.1898815. URL: https://doi.org/10.1080/0031322X.2021.1898815.

Fleiss, Joseph L. (1971). 'Measuring nominal scale agreement among many raters.' In: pp. 378–382.

Founta, Antigoni-Maria, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos and Nicolas Kourtellis (Feb. 2018). 'Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior'. In.

Gao, Lei and Ruihong Huang (2017). 'Detecting Online Hate Speech Using Context Aware Models'. In: *CoRR* abs/1710.07395. arXiv: 1710.07395. URL: http://arxiv.org/abs/1710.07395.

Gibert, Ona de, Naiara Perez, Aitor García-Pablos and Montse Cuadros (Oct. 2018). 'Hate Speech Dataset from a White Supremacy Forum'. In: *Proceedings of the 2nd Workshop*

on Abusive Language Online (ALW2). Brussels, Belgium: Association for Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/W18-5102. URL: https://www.aclweb.org/anthology/W18-5102.

Glavaš, Goran, Mladen Karan and Ivan Vulić (Dec. 2020). 'XHate-999: Analyzing and Detecting Abusive Language Across Domains and Languages'. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6350–6365. DOI: 10.18653/v1/2020.coling-main.559. URL: https://aclanthology.org/2020.coling-main.559.

Golbeck, Jennifer, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan and Derek Michael Wu (2017). 'A Large Labeled Corpus for Online Harassment Research'. In: *Proceedings of the 2017 ACM on Web Science Conference*. WebSci '17. Troy, New York, USA: Association for Computing Machinery, pp. 229–233. ISBN: 9781450348966. DOI: 10.1145/3091478.3091509. URL: https://doi.org/10.1145/3091478.3091509.

Goldman, Josephine (Dec. 2020). 'Can Black Lives Matter in a Race-Blind France? French Avoidance of 'Race' and Mobilisation of Black Collective Identity in Response to Police Brutality'. In.

Hassan, Ashif (2018). 'Hate Crime in Europe: Focusing on France & Ireland'. In: URL: https://www.academia.edu/37878580/Hate_Crime_in_Europe_Focusing_on_France_and_Ireland.

Jaki, Sylvia and Tom De Smedt (2019). 'Right-wing German Hate Speech on Twitter: Analysis and Automatic Detection'. In: *CoRR* abs/1910.07518. arXiv: 1910.07518. URL: http://arxiv.org/abs/1910.07518.

Karunanayake, Yohan, Uthayasanker Thayasivam and Surangika Ranathunga (July 2019). 'Transfer Learning Based Free-Form Speech Command Classification for Low-Resource Languages'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 288–294. DOI: 10.18653/v1/P19-2040. URL: https://aclanthology.org/P19-2040.

Kennedy, Brendan, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Koombs, Shreya Havaldar, G J Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Olmos, Adam Omary, Christina Park, Clarisa Wang, Xin Wang and Morteza Dehghani (Feb. 2020). 'The Gab Hate Corpus: A collection of 27k posts annotated for hate speech'. In: DOI: 10.31234/osf.io/hqjxn.

Kennedy, Brendan, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani and Xiang Ren (2020). 'Contextualizing Hate Speech Classifiers with Post-hoc Explanation'. In: *CoRR* abs/2005.02439. arXiv: 2005.02439. URL: https://arxiv.org/abs/2005.02439.

Khan, Muhammad Moin, Khurram Shahzad and Muhammad Kamran Malik (Mar. 2021). 'Hate Speech Detection in Roman Urdu'. In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 20.1. ISSN: 2375-4699. DOI: 10.1145/3414524. URL: https://doi.org/10.1145/3414524.

Kilvington, Daniel (2021). 'The virtual stages of hate: Using Goffman's work to conceptualise the motivations for online hate'. In: *Media, Culture & Society* 43.2, pp. 256–272. DOI:

10.1177/0163443720972318. eprint: https://doi.org/10.1177/0163443720972318. URL: https://doi.org/10.1177/0163443720972318.

Klenner, Manfred (2018). 'Offensive language without offensive words (OLWOW)'. In: *KONVENS, Germeval Task 2018 — Shared Task on the Identification of Offensive Language, Wien*. URL: https://www.zora.uzh.ch/id/eprint/159174/.

Loshchilov, Ilya and Frank Hutter (2017). 'Fixing Weight Decay Regularization in Adam'. In: *CoRR* abs/1711.05101. arXiv: 1711.05101. URL: http://arxiv.org/abs/1711.05101.

Maria Constantinou, Fabienne Baider et (Nov. 2021). 'Discours de haine dissimulée, discours alternatifs et contre-discours'. In: DOI: 10.4000/semen.12275. URL: http://journals.openedition.org/semen/12275.

Mathew, Binny, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal and Animesh Mukherjee (May 2021). 'HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection'. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.17, pp. 14867–14875. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17745.

Mozafari, Marzieh, Reza Farahbakhsh and Noël Crespi (Aug. 2020). 'Hate speech detection and racial bias mitigation in social media based on BERT model'. In: *PLOS ONE* 15.8, pp. 1–26. DOI: 10.1371/journal.pone.0237861. URL: https://doi.org/10.1371/journal.pone.0237861.

Njagi, Dennis, Z. Zuping, Damien Hanyurwimfura and Jun Long (Apr. 2015). 'A Lexicon-based Approach for Hate Speech Detection'. In: *International Journal of Multimedia and Ubiquitous Engineering* 10, pp. 215–230. DOI: 10.14257/ijmue.2015.10.4.21.

Nockleby, John T. (2000). 'Hate Speech'. In: ed. by Leonard W. Levy, Kenneth L. Karst and Dennis J. Mahoney. Encyclopedia of the American Constitution, Macmillan, 2nd edition, pp. 1277–1279.

Ousidhoum, Nedjma, Zizheng Lin, Hongming Zhang, Yangqiu Song and Dit-Yan Yeung (Nov. 2019a). '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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4675–4684. DOI: 10.18653/v1/D19-1474. URL: https://aclanthology.org/D19-1474.

— (2019b). 'Multilingual and Multi-Aspect Hate Speech Analysis'. In: *CoRR* abs/1908.11049. arXiv: 1908.11049. URL: http://arxiv.org/abs/1908.11049.

Pesole, Annarosa, Cesira Urzi Brancati, Enrique Macias, González I. and Federico Biagi (June 2018). 'Platform Workers in Europe Evidence from the COLLEEM Survey'. In: DOI: 10.2760/742789.

Pitenis, Zeses, Marcos Zampieri and Tharindu Ranasinghe (2020). 'Offensive Language Identification in Greek'. In: *CoRR* abs/2003.07459. arXiv: 2003.07459. URL: https://arxiv.org/abs/2003.07459.

Pohjonen, Matti (July 2019). 'A Comparative Approach to Social Media Extreme Speech: Online Hate Speech as Media Commentary'. In: *International Journal of Communication* 13, pp. 3088–3103.

Pohjonen, Matti and Sahana Udupa (Jan. 2017). 'Extreme Speech Online: An Anthropological Critique of Hate Speech Debates'. In: *International Journal of Communication* 11, pp. 1173–1191.

Priniski, J. Hunter, Negar Mokhberian, Bahareh Harandizadeh, Fred Morstatter, Kristina Lerman, Hongjing Lu and P. Jeffrey Brantingham (2021). 'Mapping Moral Valence of Tweets Following the Killing of George Floyd'. In: *CoRR* abs/2104.09578. arXiv: 2104.09578. URL: https://arxiv.org/abs/2104.09578.

Raha, Tathagata, Ishan Sanjeev Upadhyay, Radhika Mamidi and Vasudeva Varma (Aug. 2021). 'IIITH at SemEval-2021 Task 7: Leveraging transformer-based humourous and offensive text detection architectures using lexical and hurtlex features and task adaptive pretraining'. In: *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Online: Association for Computational Linguistics, pp. 1221–1225. DOI: 10.18653/v1/2021.semeval-1.173. URL: https://aclanthology.org/2021.semeval-1.173.

Rezvan, Mohammadreza, Saeedeh Shekarpour, Lakshika Balasuriya, Krishnaprasad Thirunarayan, Valerie L. Shalin and Amit P. Sheth (2018). 'Publishing a Quality Context-aware Annotated Corpus and Lexicon for Harassment Research'. In: *CoRR* abs/1802.09416. arXiv: 1802.09416. URL: http://arxiv.org/abs/1802.09416.

Rizoiu, Marian-Andrei, Tianyu Wang, Gabriela Ferraro and Hanna Suominen (2019). 'Transfer Learning for Hate Speech Detection in Social Media'. In: *CoRR* abs/1906.03829. arXiv: 1906.03829. URL: http://arxiv.org/abs/1906.03829.

Ross, Björn, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky and Michael Wojatzki (Sept. 2016). 'Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis'. In: *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*. Ed. by Michael Beißwenger, Michael Wojatzki and Torsten Zesch. Vol. 17. Bochumer Linguistische Arbeitsberichte. Bochum, pp. 6–9.

Roy, Sayar Ghosh, Ujwal Narayan, Tathagata Raha, Zubair Abid and Vasudeva Varma (2021). 'Leveraging Multilingual Transformers for Hate Speech Detection'. In: *CoRR* abs/2101.03207. arXiv: 2101.03207. URL: https://arxiv.org/abs/2101.03207.

Salminen, Joni, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi and Bernard J. Jansen (2020). 'Developing an online hate classifier for multiple social media platforms'. In: *Hum. Cent. Comput. Inf. Sci.* URL: https://doi.org/10.1186/s13673-019-0205-6.

Salminen, Joni, Fabio Veronesi, Hind Almerekhi, Soon-Gvo Jung and Bernard J. Jansen (2018). 'Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings'. In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 88–94. DOI: 10.1109/SNAMS.2018.8554954.

Sap, Maarten, Dallas Card, Saadia Gabriel, Choi Yejin and Noah Smith (Jan. 2019). 'The Risk of Racial Bias in Hate Speech Detection'. In: pp. 1668–1678. DOI: 10.18653/v1/P19-1163.

Sap, Maarten, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith and Yejin Choi (July 2020). 'Social Bias Frames: Reasoning about Social and Power Implications of Language'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5477–5490. DOI: 10.18653/v1/2020.acl-main.486. URL: https://aclanthology.org/2020.acl-main.486.

Schmidt, Anna and Michael Wiegand (Jan. 2017). 'A Survey on Hate Speech Detection using Natural Language Processing'. In: pp. 1–10. DOI: 10.18653/v1/W17-1101.

Schröder, Janik (2020). 'Entwicklung eines Browser-Plugins zur nutzerseitigen Filtration von Hate Speech in sozialen Netzwerken'. MA thesis. Universität Hamburg.

Seoane, Annabelle and Angeliki Monnier (2019). 'Discours de haine sur l'internet.' In: hal: hal-02153771. URL: https://hal.archives-ouvertes.fr/hal-02153771/document.

Siapera, Eugenia (Feb. 2019). 'Organised and Ambient Digital Racism: Multidirectional Flows in the Irish Digital Sphere'. In: *Open Library of Humanities* 5. DOI: 10.16995/olh.405.

Thelwall, Michael and Saheeda Thelwall (2021). 'Twitter during COVID-19: George Floyd Opening a Space to Address Systematic and Institutionalized Racism?' In: DOI: 10.1177/0163443720972318. URL: https://ssrn.com/abstract=3764867.

Tontodimamma, Alice, Eugenia Nissi, Annalina Sarra and Lara Fontanella (2021). 'Thirty years of research into hate speech: topics of interest and their evolution.' In: *Scientometrics 126*. DOI: 10.1007/s11192-020-03737-6.

Tulkens, Stéphan, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven and Walter Daelemans (2016). 'A Dictionary-based Approach to Racism Detection in Dutch Social Media'. In: *CoRR* abs/1608.08738. arXiv: 1608.08738. URL: http://arxiv.org/abs/1608.08738.

Vanetik, Natalia and Elisheva Mimoun (2022). 'Detection of Racist Language in French Tweets'. In: *Information* 13.7. ISSN: 2078-2489. DOI: 10.3390/info13070318. URL: https://www.mdpi.com/2078-2489/13/7/318.

Vidgen, Bertie and Leon Derczynski (2020). 'Directions in Abusive Language Training Data: Garbage In, Garbage Out'. In: *CoRR* abs/2004.01670. arXiv: 2004.01670. URL: https://arxiv.org/abs/2004.01670.

Vogel, Inna, Roey Regev and Martin Steinebach (2019). 'Automatisierte Analyse Radikaler Inhalte im Internet'. In: *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik – Informatik für Gesellschaft*. Ed. by Klaus David, Kurt Geihs, Martin Lange and Gerd Stumme. Bonn: Gesellschaft für Informatik e.V., pp. 233–245. DOI: 10.18420/inf2019_27.

Wang, Simeng, Xiabing Chen, Yong Li, Chloé Luu, Ran Yan and Francesco Madrisotti (2021). ''I'm more afraid of racism than of the virus!': racism awareness and resistance among Chinese migrants and their descendants in France during the Covid-19 pandemic'. In: *European Societies* 23.sup1, S721–S742. DOI: 10.1080/14616696.2020.1836384. eprint: https://doi.org/10.1080/14616696.2020.1836384. URL: https://doi.org/10.1080/14616696.2020.1836384.

Warner, William and Julia Hirschberg (June 2012). 'Detecting Hate Speech on the World Wide Web'. In: *Proceedings of the Second Workshop on Language in Social Media*. Montréal, Canada: Association for Computational Linguistics, pp. 19–26. URL: https://aclanthology.org/W12-2103.

Waseem, Zeerak (Nov. 2016). 'Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter'. In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Austin, Texas: Association for Computational Linguistics, pp. 138–142. DOI: 10.18653/v1/W16-5618. URL: https://aclanthology.org/W16-5618.

Waseem, Zeerak and Dirk Hovy (Jan. 2016). 'Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter'. In: pp. 88–93. DOI: 10.18653/v1/N16-2013.

Yimam, Seid Muhie, Abinew Ali Ayele and Chris Biemann (2019). 'Analysis of the Ethiopic Twitter Dataset for Abusive Speech in Amharic'. In: *Proceedings of International Conference On Language Technologies For All: Enabling Linguistic Diversity And Multilingualism Worldwide (LT4ALL 2019)* abs/1912.04419. arXiv: 1912.04419. URL: https://www.inf.uni-hamburg.de/en/inst/ab/lt/publications/2019-yimametal-lt4all-amharichate.pdf.

Yin, Wenjie and Arkaitz Zubiaga (June 2021). 'Towards generalisable hate speech detection: a review on obstacles and solutions'. In: *PeerJ Computer Science* 7, e598. DOI: 10.7717/peerj-cs.598.

Zampieri, Marcos, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra and Ritesh Kumar (June 2019). 'Predicting the Type and Target of Offensive Posts in Social Media'.

In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 1415–1420. DOI: 10.18653/v1/N19-1144. URL: https://aclanthology.org/N19-1144.

A. Annotation Guidelines

Racial hate speech detection

The data will be used for academic research to advance the hate speech detection, especially for languages other than English. The contents do not reflect the views of the requesters. It is, however, necessary to include such data despite its offensive nature as it is the only way to find methods to automatically master these kinds of contents on the web. Thank you very much for your contribution to the research! You can optionally indicate your gender and ethnicity, this would help us a lot.

Rules

Performers who speak English will get access to the tasks. There is a short training session such that you can get used to the tasks.

Classification

When you start a task suite, a text will appear which you read and classify into one of the 4 options: hate speech, offensive speech, normal speech or unsure if you are unsure. If the text contains hate speech or offensive speech, please specify if the content is direct against a target group or person based on racism. You will find definitions of the terms below. Please consider if the text is generally hateful or offensive and avoid your personal perception (e.g. if you are offended by the text). Do not base your classification only on the words but rather on the context of the whole text.

You can pause any time, especially if you notice signs of negative stress or emotions. If you need help, you can contact the helpline on https://www.crisistextline.org/. Please respect the privacy of the content creators. We might reject your work and ban you in case of patterns of random labeling. Warning: Please note that the data contains hateful or sensitive content, including racial slurs.

Glossary

Hate speech is discriminating attack or text directed towards a person or group based on a specific characteristics like race, color, ethnicity, national origin, nationality, gender (identity), sexual orientation, sex, religion, caste, age, disability, or serious disease. This/These characteristic(s) needs to be the cause for the threat. It can contain violence or threats and usually has the intention to wound. Offensive speech is speech that causes anger, contains non-acceptable language (profanity) or that is hurtful, it can be a personal attack or a general insult. It might contain swear words, slurs or support problematic hash tags like "#BanIslam" or "#whitegenocide".

Normal speech is speech that does not fit in either category and can be considered as non-problematic language. Racism includes all negative utterances, negative generalizations and insults concerning ethnicity, nationality, religion and culture.

Example classifications

#CelebrityLeakedPhotos YEAH NOW YOU KNOW how it feels knowing the#NSA is on all our phones. Quit whining like#muslim#bitches. offensive speech Jews and white race traitors have to be swept clean so the Aryan race can survive. hate speech, racial target Today, do something for the environment and kill your local queer. hate speech, non-racial target A nigga been going through so much stress and grief this year and not a soul eased up on asking me for favors this year, life fuckin crazy. normal speech

Rassistische Hatespeecherkennung

Die Daten aus dieser Aufgabe werden für die Forschung benutzt, um Hatespeecherkennung von anderen Sprachen als Englisch voran zu treiben. Der Inhalt der Tweets spiegelt nicht unsere Meinung wieder. Allerdings ist es nötig, Hassrede manuell zu klassifizieren, um diese später automatisch im Internet erkennen zu können. Wir bedanken uns herzlich für Ihre Teilnahme! Es hilft uns sehr, wenn Sie zusätzlich noch ihr Geschlecht und ihre Ethnie(n) angeben.

Regeln

Performer, die Deutsch sprechen, bekommen Zugang zu unseren Klassifikationsaufgaben. Es gibt einige Trainingsaufgaben, um sich mit der Art der Aufgaben vertraut zu machen.

Klassifizierung

Wenn sie eine Task Suite öffnen, wird pro Task ein kleiner Text, ein Tweet, erscheinen. Bitte klassifizieren Sie diesen in eine der vier Optionen: hate speech (Hassrede), offensive (Beleidigung), normal oder unsure (unsicher), wenn Sie sich nicht sicher sind, welche der vorigen Optionen am besten passt. Enthält der Text Hatespeech oder Beleidigungen, geben Sie bitte an, ob diese rassistisch sind. Sie finden Definitionen der Begriffe weiter unten. Bitte orientieren Sie sich bei der Klassifizierung daran, ob der Text allgemein beleidigend/hasserfüllt ist und nicht, ob Sie sich davon angegriffen fühlen. Betrachten Sie den gesamten Kontext und nicht nur einzelne Wörter.

Sie können jederzeit unterbrechen, besonders bei Anzeichen von negativem Stress or Gefühlen. Wenn Sie Hilfe benötigen, können Sie die englischsprachige, chatbasierte Hilfehotline kontaktieren: https://www.crisistextline.org/.

Bitte respektieren Sie die Privatsphäre der Verfasser der Tweets. Wir behalten uns vor, Ihre Bearbeitung der Aufgaben abzulehnen und Sie zu blockieren, sollten wir Anzeichen für zufällige Klassifizierung bemerken.

Warnung: Die Tweets können teilweise hasserfüllte oder sensible Inhalte wie rassistische Beleidigungen enthalten.

Glossar

Hate speech ist ein diskriminierender Angriff oder Text gegen eine Person oder Gruppe, basierend auf bestimmten Eigenschaften. Diese können die Race, Hautfarbe, (zugeschriebene) Ethnie, Herkunft, Staatsbürgerschaft, Gender(-Identität), biologisches Geschlecht, Religion, Kaste, Behinderungen oder eine schwere Krankheit sein. Diese Eigenschaft(en) muss/müssen die Ursache für die Beleidigung/den Angriff sein. Diese kann Gewalt, Bedrohungen enthalten und hat üblicherweise die Absicht, zu verletzen. Offensive/ Beleidigungen verletzen, verursachen Ärger oder enthalten nicht-akzeptierte Sprache wie Obszönität. Diese können sowohl persönlich sein als auch allgemein beleidigend. Schimpfwörter, Verleumndungen oder problematische Hashtags wie "#BanIslam" oder "#whitegenocide" können enthalten sein. Normale Sprache passt in keine der beiden oberen Kategorien und kann als nicht problematisch eingestuft werden. Racism/ Rassismus schließt alle negativen Äußerungen, Verallgemeinerungen and Beleidigungen, die sich gegen die Ethnie, die Nationalität, die Religion oder die Kultur einer Person(engruppe) richten.

Beispielklassifizierungen

Beispiele aus Studie von Ross et al.[1]

- Warum lädt man eigentlich nicht mal einen normalen Bürger zu#Maischberger ein, wenn es um das Thema#Asylanten geht?. offensive (Beleidigung)
- #maischberger Liefers: Wir achten Frauen.... Für viele von den#rapefugees ist das wie ein Paradies hier.Lasche Strafen bei Vergewaltigungen. hate speech, racial target Hatespeech, non-racial target (Hassrede, rassistisch)
- bitte nicht die#Türkei zum#EU-Mitglied machen!#Menschenrechte#Pressefreiheit#Islamisierung. Normal speech (Normale Sprache)

Détection des discours de haine raciale

On va utiliser les données pour la recherche académique pour avancer la détection des discours de haine en Français. Les contenus des texts ne reflètent pas l'opinion des demandeurs. Pour avancer la détection automatique on doit les intégrer et les classer manuellement. Merci beaucoup pour votre contribution! Vous avez la possibilité de nous indiquer votre genre et ethnique pour nous aider en plus.

Règlement

Les performers qui parlent fraînçais peuvent accéder notre tâches. Il y a quelques tâches de formation ibt einige Trainingsaufgaben, pour se familiariser avec les tâches.

Classification

Quand vous commencez une suite de tâches, vouz verrez un text pour le classer. Il y a quatre catégories: hate speech (discours de haine), offensive speech (insulte), normal speech (normale)

[1] https://github.com/UCSM-DUE/IWG_hatespeech_public/blob/master/german%20hatespeech%20refugees.csv

ou unsure (incertain) si vous êtes incertain. Si le text est la discours de hain ou un insulte, indiquez si les contenu ont un objectif racial contre une personne ou groupe. Vouz trouverez des définitions au dessous. Ne fondez pas votre classification sur votre perception personnelle, mais plutôt sur le fait qu'il s'agisse d'un acte généralement haineux. Fondez votre classification sur le contexte et pas seulement sur un seule mot.

Vous pouvez faire une pause à tout moment, notamment si vous remarquez des signes de stress ou d'émotions négatives. Si vous avez besoin d'écrire avec un professionaliste, vous pouvez contacter le service d'assistance (en anglais) https://www.crisistextline.org/.

Veuillez respecter la vie privée des créateurs de contenu. Nous pourrions rejeter votre travail et vous bannir en cas d'étiquetage aléatoire. Avertissement : Veuillez noter que ces données contiennent des contenus haineux ou sensibles, notamment des insultes raciales.

Glossaire

Hate speech (discours de haine) est une attaque ou un text discriminatoire contre une personne ou un groupe basé sur un propriété spécifique comme l'ethnie, la race, lórigine nationale, le genre (identité), l'orientation sexuelle, le sexe, la caste, l'âge, le handicap, la religion ou une maladie grave. Le(s) characteristique(s) doivent être la cause pour la menace. Il(s) peut/peuvent comprendre de la violence ou de menaces et normalement a l'intention de blesser. Offensive speech (insulte) est un discours qui provoque la colère, contient un langage inacceptable (blasphème) ou qui est blessant, il peut s'agir d'une attaque personnelle ou d'une insulte générale. Il peut contenir des jurons, des insultes ou des tags problématiques tels que "#BanIslam" ou "#whitegenocide". Normal speech est un discours qui n'entre dans aucune des deux catégories et qui peut être considéré comme un langage non problématique. Racism comprend tous les propos négatifs, les généralisations négatives et les insultes concernant l'ethnicité, la nationalité, la religion et la culture.

Exemples de classifications

- Ma femme est hystorique. C'est comme hystérique, sauf que losqu'elle et un cable elle me sort des vieux dossiers. offensive speech (insulte)
- Nous sommes tous racistes parce que nous ne voulons pas la destruction de l'identité culturelle de notre pays. hate speech, racial target (discours de haine raciale) [2]
- les femmes qui sont en plus Dijonnaise ne parlez pas de foot sivouplai c'est comme si un aveugle manchot parler de passer le permis. hate speech, non-racial target (discours de hain non raciale) (Chiril, Moriceau et al. 2020)
- Laetitia Casta pas d'accord avec#balancet. normal (discours normal) (Chiril, Moriceau et al. 2020)

[2] https://www.cairn.info/revue-deviance-et-societe-2019-3-page-359.htm

List of Figures

Eidesstattliche Erklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Masterstudiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Hamburg, den 4.10.2022 Skadi Dinter

Veröffentlichung

Ich stimme der Einstellung der Arbeit in die Bibliothek des Fachbereichs Informatik zu.

Hamburg, den 4.10.2022 Skadi Dinter