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

Master Thesis - Oral Defence

Frederik Gaasdal Jensen 141628
Henry Alexander Stoll 141636

Supervisor: Raghava Rao Mukkamala
Co-Supervisor: Sippo Rossi

CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Overview

Agenda

1. Introduction
2. Methodology
3. Results
4. Refugee Crises - Prediction
5. Discussion/Conclusion

Introduction

Automatically detecting hate speech surrounding refugees on social media

UNHCR, the UN Refugee agency is mandated by the United Nations to lead and coordinate international action for the worldwide protection of refugees and the resolution of refugee problems.

- 84 million people forcibly displaced, **26.6 million refugees**[1]
- Hate speech and incitement of violence on social media platforms are **linked to** acts of physical **violence**
- This includes atrocity crimes like the **genocide** against Rohingya Muslims forcing 700.000 people to flee
- UNHCR is tasked with **monitoring online hate speech** against refugees, which is being done **manually** to date

Key Challenges detecting hate speech

- Classifying hate speech is **complex and nuanced**, difficult for machines, humans, and experts
- Large variety of definitions and overlap with similar terms
- **Changing cultural norms** and awareness
- Low inter-annotator agreements
- Lack of large, robust and detailed datasets
- **Limited research** on hate speech targeted at refugees
- Resource-intensive task to manually analyze social media posts on a large scale

Research Question

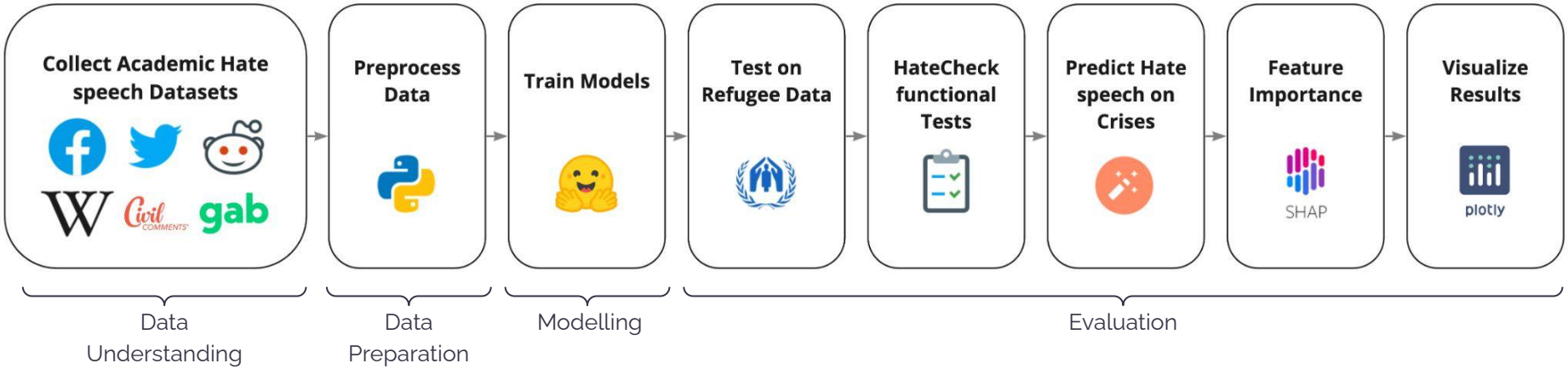
*How can **hate speech against refugees** on social media platforms be **detected** and **measured** using Natural Language Processing methods?*

- In general and in the context of refugees
- Most important features
- Development over time

[1]as of mid-2021

Methodology

The overall methodological process consists of 8 different phases



Collected 12 labeled hate speech datasets from various contexts...

Selection Criteria

Origin, size, actuality, data collection strategy, annotation scheme, guidelines and process, as well as inter-annotator agreement were all considered as factors.

Source

Twitter, Reddit, Facebook, Gab, Civil Comments, Wikipedia, adversarially generated, collected between 2017-2021 with examples dating back to 2010

Collection Strategy

Seed-based, community-based, synthetic data from trained annotators, hate speech or slur lexicon-based, random sampling

Annotation

Wide variety of annotation schemes: binary, tertiary, multi-class, multi-level (targets and types of hate speech) with little consistency

Size

5,000 rows up to 2,000,000 rows of annotated data

Preprocessing

Lowercasing sample, removing user mentions, URLs, hashtags, converting emojis to standard text representation

Label Mapping

Standardizing all annotation schemes to converge into three classes: hate speech, offensive and normal for everything else

	Normal	Offensive	Hate Speech	Total
N Samples	263,406	135,190	74,023	472,619
Avg Words Per Sample	42	36	27	38
Word Count	11,152,326	4,859,976	1,967,928	17,980,230

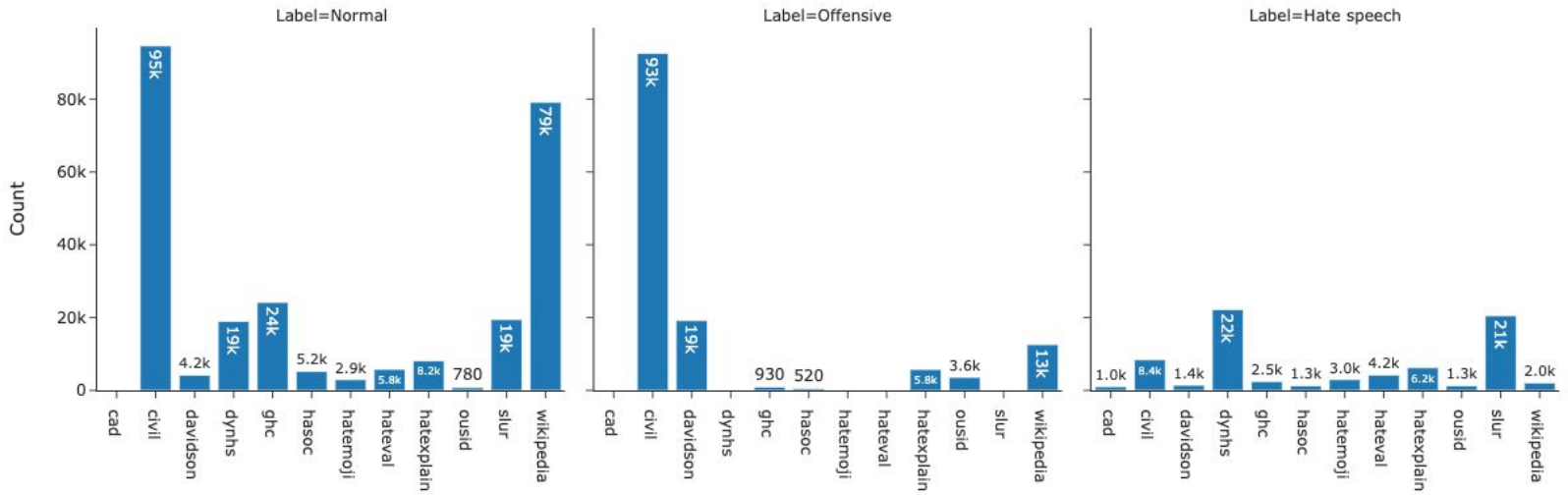
...to create one combined dataset with 470k samples for model training including variations

Data Split

80/10/10 for train, eval, and test data.

Oversampling

Based on success in previous research a dataset variation was created where hate speech classes was oversampled.



To test deep learning models' ability to detect hate speech surrounding refugees, additional datasets were obtained

UNHCR - Hate Speech Dataset

- 98 tweets with hate speech
- A few tweets were translated into English by UNHCR
- Annotated by a single employee

> *"#RefugeesWelcome is a terrorist organisation that is helping BORIS's government with their Great Replacement Policy to replace all British people with the scum of the earth."*
>
> *"@USER @USER Most are not refugees, they are bogus refugees, economic migrants looking for a better life, some are criminals and some are terrorists."*

International Refugee Crisis Datasets

- **Afghanistan**, United Kingdom Channel Crossings, **Greece-Turkey**, Rohingya, Tigray
- Unlabeled Twitter data that spans over 1-2 months for each crisis
- Centered around trigger events
  - Afghanistan → Fall of Kabul
  - Greece-Turkey → Opening of the border

Crisis	Period	Number of Tweets
Afghanistan	26-07-2021 to 31-08-2021	283k
UK Channel Crossings	19-07-2020 to 29-08-2020	173k
Greece-Turkey	11-02-2020 to 23-03-2020	137k
Rohingya	01-03-2021 to 30-04-2021	29k
Tigray	15-01-2021 to 30-04-2021	42k

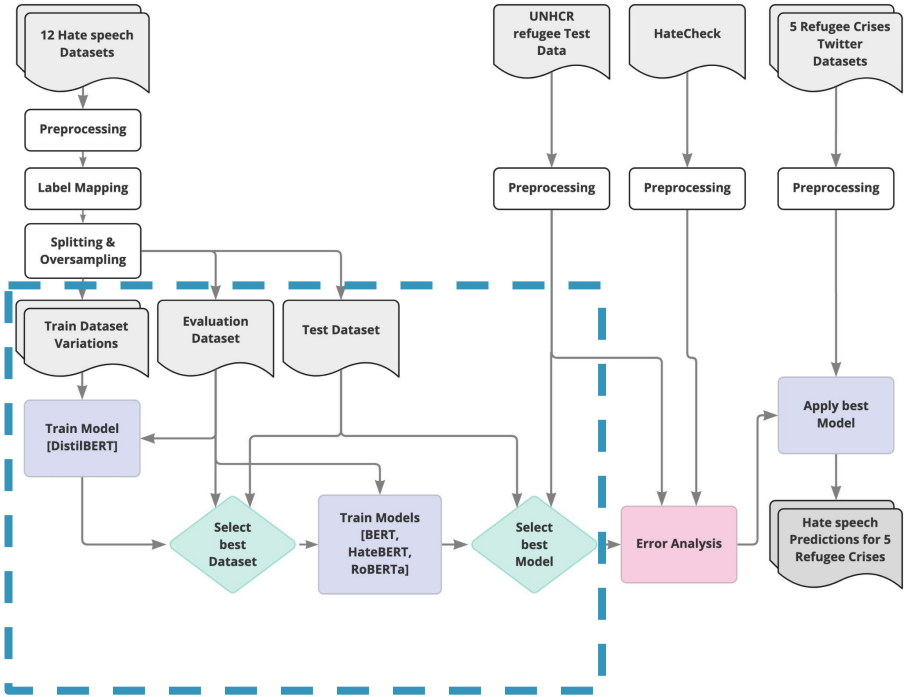
The experiments are conducted in 2 iterations

Iteration 1: Select best dataset variation

- **DistilBERT**
  - Constitutes 60% of the size of BERT
  - 60% faster and retains 97% of BERT's language understanding capabilities
  - Less expensive
- The oversampled version of the combined training dataset increased the performance

Iteration 2: Select best model

- **BERT (baseline)**
  - Base version
  - Most standard BERT model
- **HateBERT**
  - Additionally pretrained a large scale Reddit dataset
- **RoBERTa**
  - Pre-trained for a longer time using much more data, larger batches, and longer sequences



Results

Results - In-Dataset

The RoBERTa model performs best on data from the same domain as the training dataset with an F1-Score of 81.0%

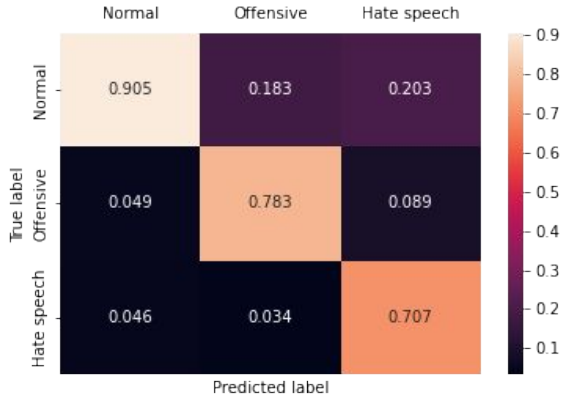
In-Dataset Evaluation:

- All models achieve a **similar F1-Score** on the same type of data as they were trained on
- The baseline model performs best on the normal class, however, the **difference is not significant**
- All models perform worst on the hate speech class.

Confusion Matrix - RoBERTa:

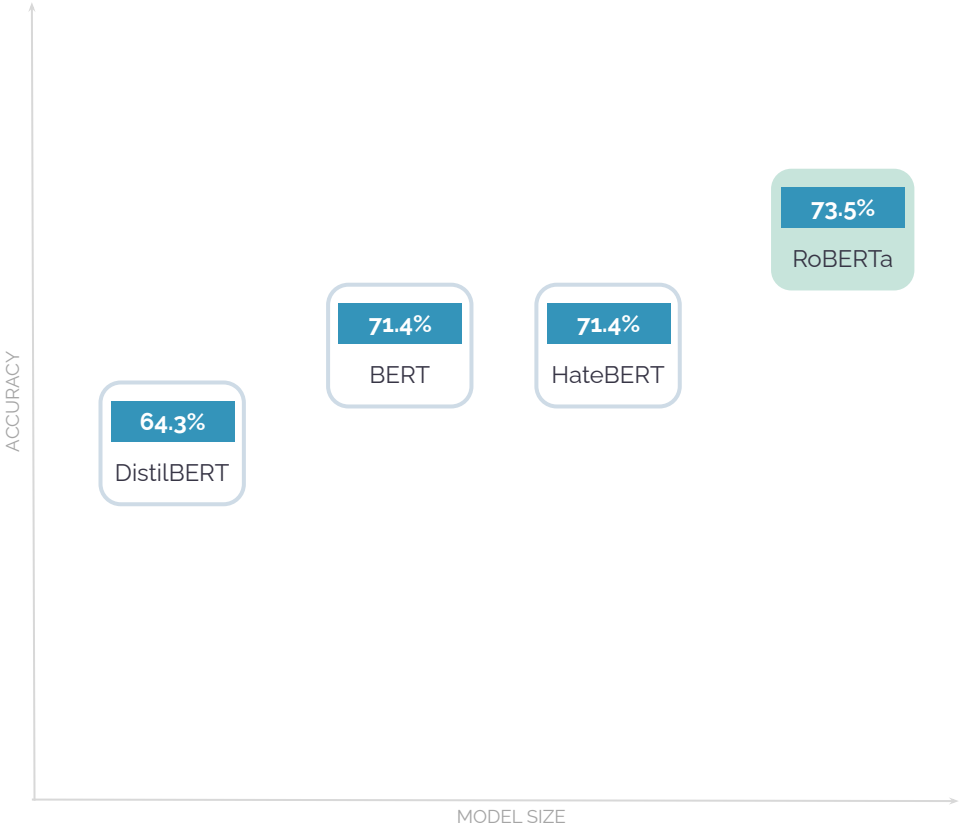
- Normalized for the predicted label
- Correctly classifies 90.5% of the normal class
- When it predicts text as hate speech, **the true class is offensive in 8.9% of the cases**
- When it predicts text as offensive , **the true class is hate speech in 3.4% of the cases**

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



Results - UNHCR Data

The RoBERTa model performs best on refugee-related data

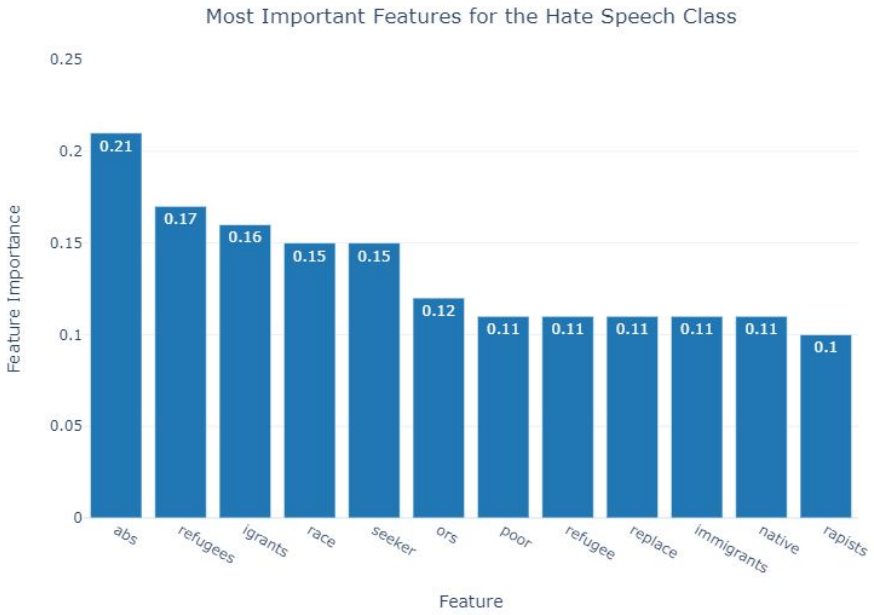


All the models are trained on the combined dataset that consists of **12 hate speech datasets** that originates from different contexts

- Indicates that multiple datasets from different contexts can be **combined to predict** hate speech **targeting refugees**
- More beneficial to use a model that is pre-trained with **more data** and for a **longer time**
- Tradeoff: **fine-tuning cost vs. accuracy**

"*abs*"[2], "*refugees*", and "*igrants*" are the most influential features for the hate speech class

- Importance score represented as the mean SHAP value
- The features have a **lower impact on average** for the UNHCR refugee-related dataset compared to the test set that was used in the in-dataset evaluation
- "*refugees*", "*igrants*", "*refugee*", and "*immigrants*" appear to be among the most influential features



[2] "arabs."

29 functional tests were applied to detect model weaknesses

	Functionality	Example Test Case	Gold Label	n	Accuracy (%) B	H	R
Derogation	**F1**: Expression of strong negative emotions (explicit)	"I hate [IDENTITY]." "You [IDENTITY] make me sick."	hateful	160	98.1	**99.4**	**99.4**
	F2: Description using very negative attributes (explicit)	"[IDENTITY] are disgusting." "All [IDENTITY] like them are vile."	hateful	160	98.1	**99.4**	98.1
	F3: Dehumanisation (explicit)	"[IDENTITY] are scum." "[IDENTITY] are a pest."	hateful	160	98.1	**99.4**	98.8
	F4: Implicit derogation	"[IDENTITY] belong in a zoo." "I'd rather die than date [IDENTITY].	hateful	160	95.6	**99.4**	97.5
Negation	**F14**: Hate expressed using negated positive statement	"[IDENTITY] lives do not matter."	hateful	160	97.5	**99.4**	**99.4**
	F15: Non-hate expressed using negated hateful statement	"No [IDENTITY] deserves to die." We shouldn't let [IDENTITY] suffer."	non-hate	152	21.1	43.4	**57.9**
Non-hate grp. ident.	**F18**: Neutral statements using protected group identifiers	We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hate	144	75.0	**77.1**	75.7
	F19: Positive statements using protected group identifiers	"I love [IDENTITY]." "[IDENTITY] are great."	non-hate	216	70.8	79.2	**87.0**
Counter Speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	190	41.6	54.7	**79.5**
	F21: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	156	62.8	51.9	**73.1**

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

The models exhibit only minor weaknesses towards refugee related groups

Overall Performance:

- RoBERTa achieves the highest **accuracy score** on both classes
- Significantly drop in performance on the non-hateful class, which indicates that the models are **struggling with non-hateful tests**
- The tests consist of difficult cases

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

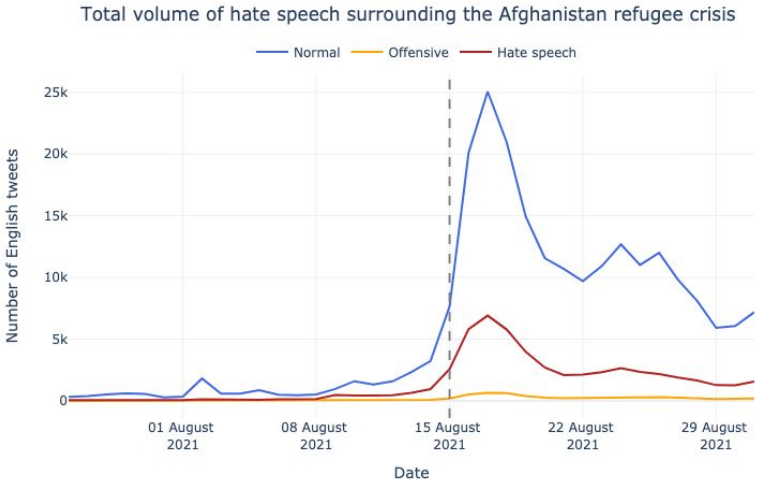
Target Group Performance:

- The HateCheck dataset was expanded with 421 test cases focusing on "Refugees" as a target group
- RoBERTa outperforms the other models on all target groups
- RoBERTa performs best on "Trans people", "Immigrants", and "Refugees"
- All models **exhibit weaknesses towards** the "**Women**" target group

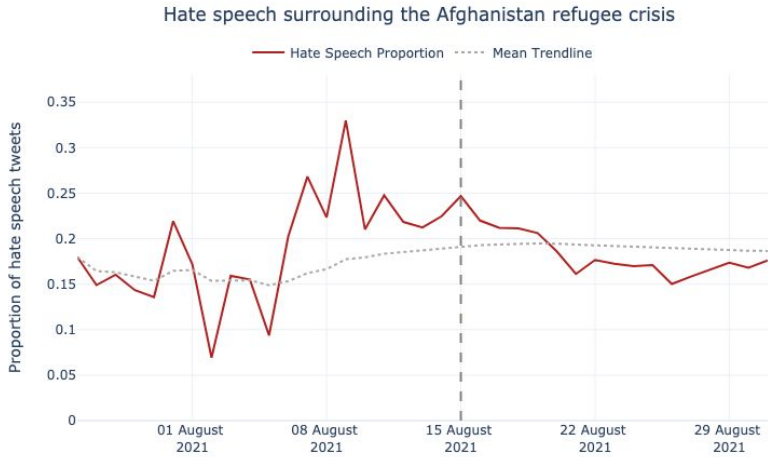
Target Group	n	BERT	HateBERT	RoBERTa
Black people	421	80.9%	84.0%	89.8%
Disabled people	421	80.4%	85.5%	90.1%
Gay people	421	81.9%	88.4%	88.9%
Immigrants	421	81.9%	88.1%	**92.2%**
Refugees	421	83.1%	89.5%	**91.0%**
Muslims	421	83.9%	87.6%	90.3%
Trans people	421	85.5%	88.3%	92.7%
Women	421	74.7%	80.2%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

Refugee Crisis Prediction

The proportion of hate speech for the Afghanistan crisis increases around trigger event but decreases afterwards

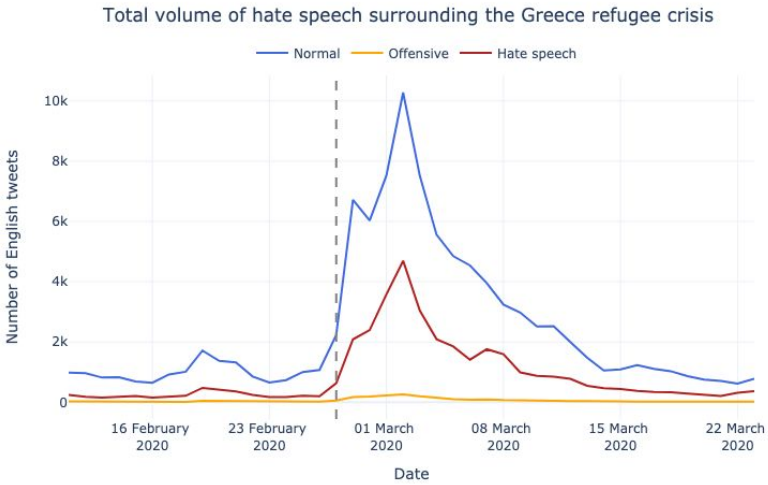


- **Low activity** level before the trigger event
- After the trigger event, the amount of hate and normal speech **increased drastically for a few days** before declining again
- The amount of offensive speech stayed consistently low

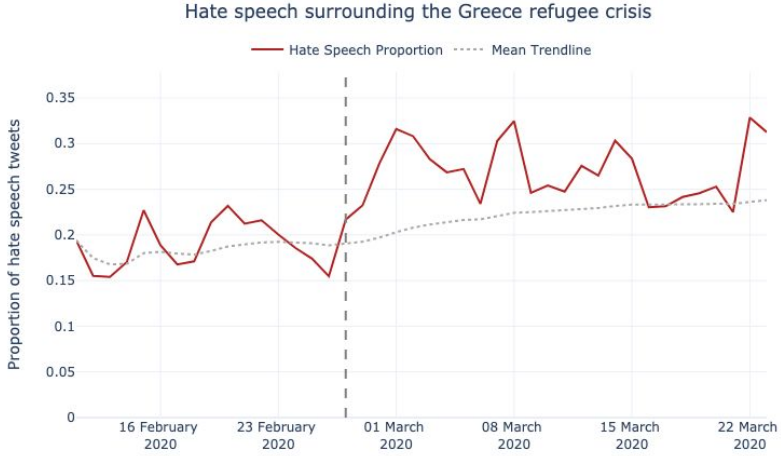


- **Ranges** between 6.9% and 32.9%
- **High fluctuation** before the trigger event (might be caused by the limited number of tweets)
- Highest proportion is on the 9th of August 2021
- The rolling average **trendline experienced a minor increase** when reaching the trigger event
- **Daily proportions decreases** from the 20th of August 2021, which affects the trendline

The proportion of hate speech for the Greece-Turkey crisis increased even though the total number of tweets decreased after the event



- The amount of hate and normal speech started to **rise on the 27th of February 2020**
- The volume experienced a **drastic increase in a 4-days period**
- The **highest activity level was reached on the 2th of March 2020**, where it started to decrease again
- Offensive speech stays consistently low throughout the whole time period



- **Fluctuates** during the whole period
- Before the trigger event, the lowest proportion of hate speech is observed
- In a **4-days period** after the trigger event, the proportion of hate speech **increased by a factor of 2** since it goes from 15.5% to 31.6%
- The rolling average trendline starts to **exhibit an increasing trend** after the trigger event

Discussion

Implementing the recommendations would increase the effectiveness and accuracy of detecting hate speech targeted refugees

Root Cause Analysis

Combining the insights from the hate speech detection system with other data sources would **increase the overall understanding** of the discussions of various crises.

Contributes to *Commitment 2* (United Nations)[3]

Monitoring System

Implementing a monitoring system makes it possible to **track the development** of hate, offensive, and normal speech on social media sites over time.

Monitoring the online discussions would make it **easier** for UNHCR to **implement preventative** measures.

Contributes to *Commitment 1* (United Nations)[3]

Research Field

Collaborating with academia would increase the **awareness** of hate speech targeted at refugees, and potentially lead to an **increase in research**.

Sharing a high quality dataset could also **stimulate more research** as the problem right now is the **lack of available data**.

[3] United Nations Strategy and Plan of Action Points regarding Hate Speech

Which limitations do the thesis possess and what is the future direction?

Limitations

Limited to English tweets - findings cannot be transferred to a different language

Combining Multiple Datasets of varying quality - different collection strategies, label definitions, and annotation strategies have been used for each of the 12 datasets

Low variety of refugee-related tweets for evaluation - **losing contextual information** by translating tweets, annotation performed by only a **single employee**

Features are context-dependent - makes it difficult to generalize

The understanding of hate speech changes over time - data needs to be reviewed in terms of annotations

Future Work

Dataset Improvements - ablation study to identify datasets that are irrelevant and hurting the performance

Add context to model - use meta-data such as profile bios, profile pictures, network information and pictures that appear in tweets

Model Improvements - the training dataset should be enriched with data from areas identified in HateCheck where the model is struggling

By implementing the suggestions for future work, we think that the performance of the hate speech detection model can be further increased.

Conclusion

Conclusion

Hate speech targeted at refugees can be detected, measured, and analyzed using state-of-the-art transformer-based BERT models

Detect

Combine and standardize general hate speech datasets

Classify hate speech with an F1 of 81.0%

Classify hate speech targeted at refugees with an accuracy of 73.5%

Show absence of major weaknesses with an HateCheck (modified) accuracy of 90.3%

Measure

Show how hate speech changes over time

In the context of specific refugee crises

Measure differences between conflicts

Enable Root Cause Analysis

Analyze

Identify which features are most important for hate and offensive speech

Compare top features of general and refugee-related datasets

Identify top features for each refugee crisis

Results were validated by UNHCR and the model will be used in production

Appendix

All the data sources have be preprocessed by applying the following 6 steps

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

All labels are standardized into three classes: Hate, Offensive, and Normal

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

The RoBERTa model misclassified 3 tweets as offensive and 23 as normal

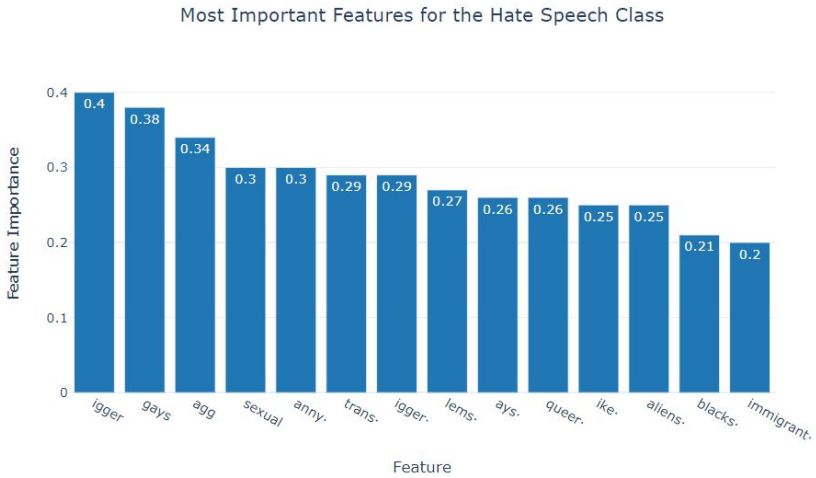
Misclassified Example	Prediction	Normal	Offensive	Hate Speech
"@USER @USER turkey welcomes more than 5 million syrian refugees, remind me when france opens its borders at only 1 million dirty puffon"	Offensive	0.15	**0.45**	0.40
"@USER @USER granting asylum to refugees fleeing from oppression who are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

Confidence Scores

“*igger*”, “*gays*”, and “*agg*”[3] are the most influential features for the hate speech class

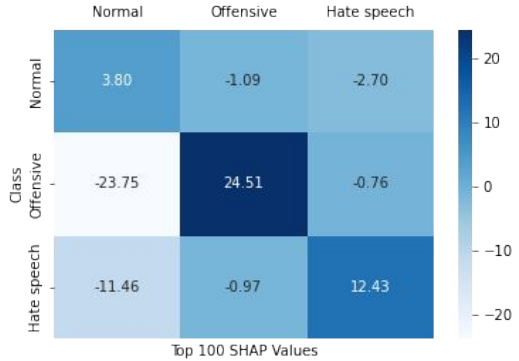
Most Important Features for the Hate Speech class:

- Tokens that end with a whitespace have been marked with “·”
- Tokens that refer to the identity of a person are the most influential ones
- Not generalizable as the tokens depend on the context they appear in



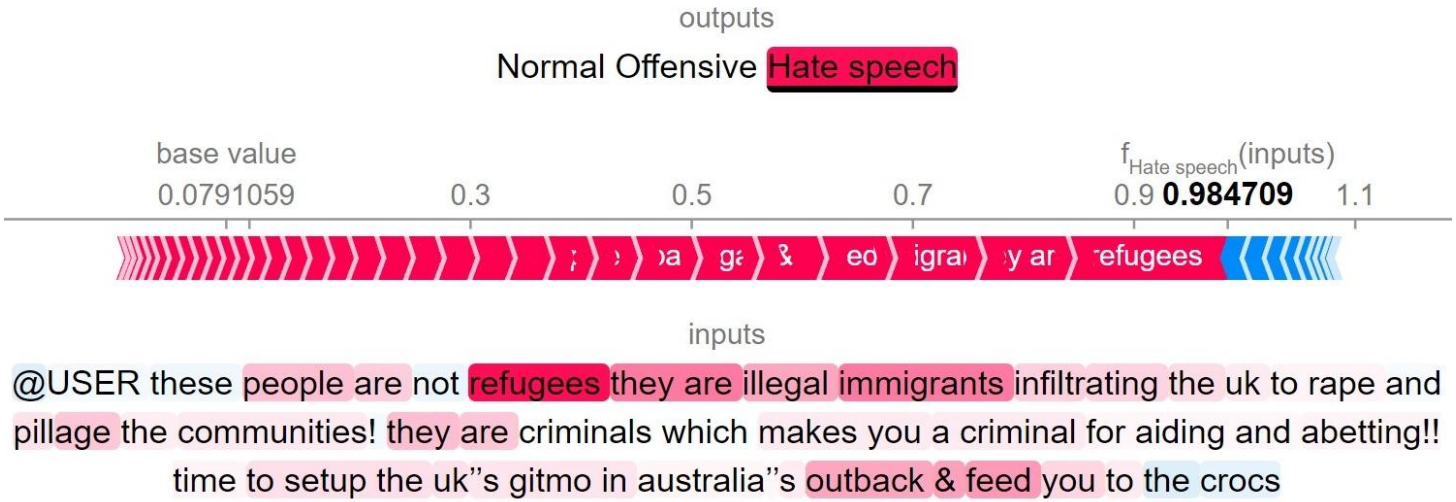
Feature Importance of Top 100 Tokens from each Class:

- Cumulative contribution
- Mean SHAP value summed up for each class
- Offensive features have twice as much impact on itself compared to the hate speech features
- Top offensive and hate speech features have little impact on each other

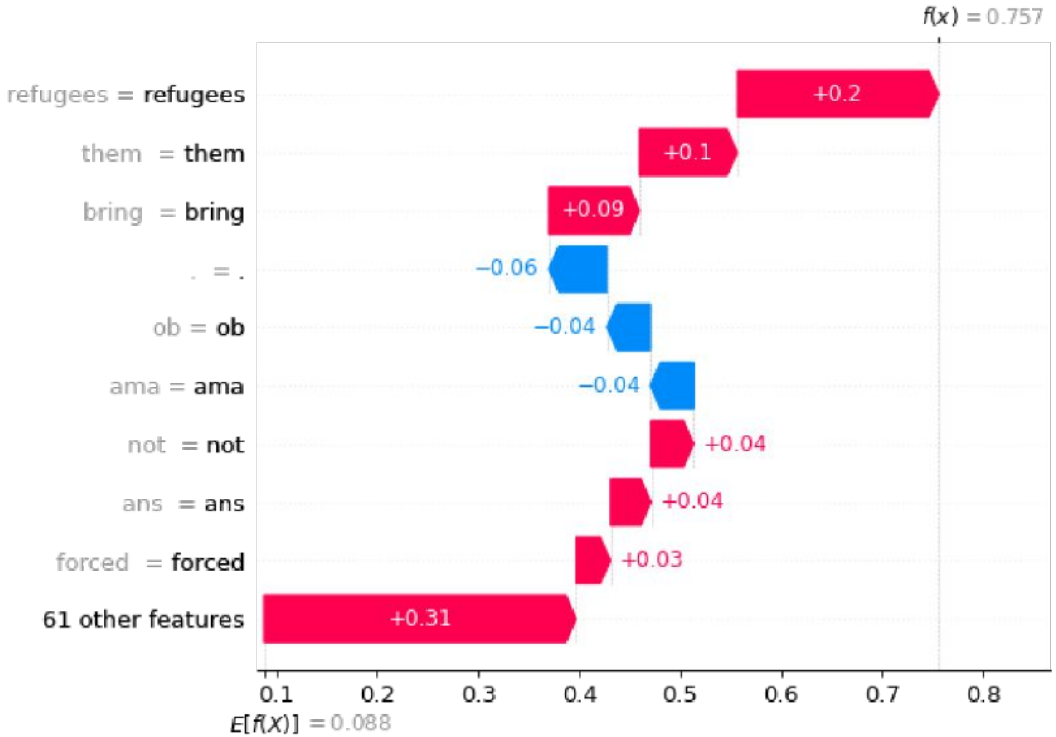


[3] "faggot...", "aggies", "teagger", "ragged", "self-aggrandizing"

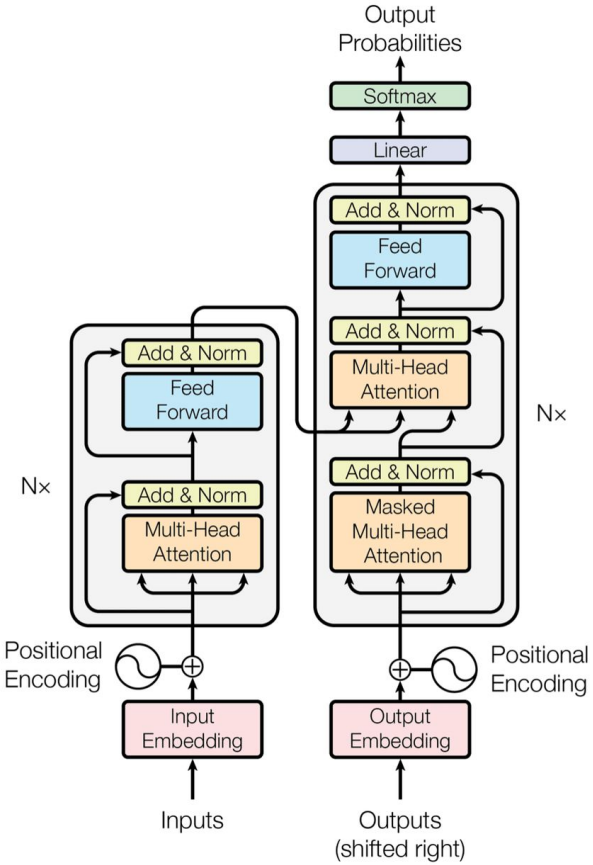
Analyzing how the model performed a prediction reveals insights about the influential features on a sentence-level



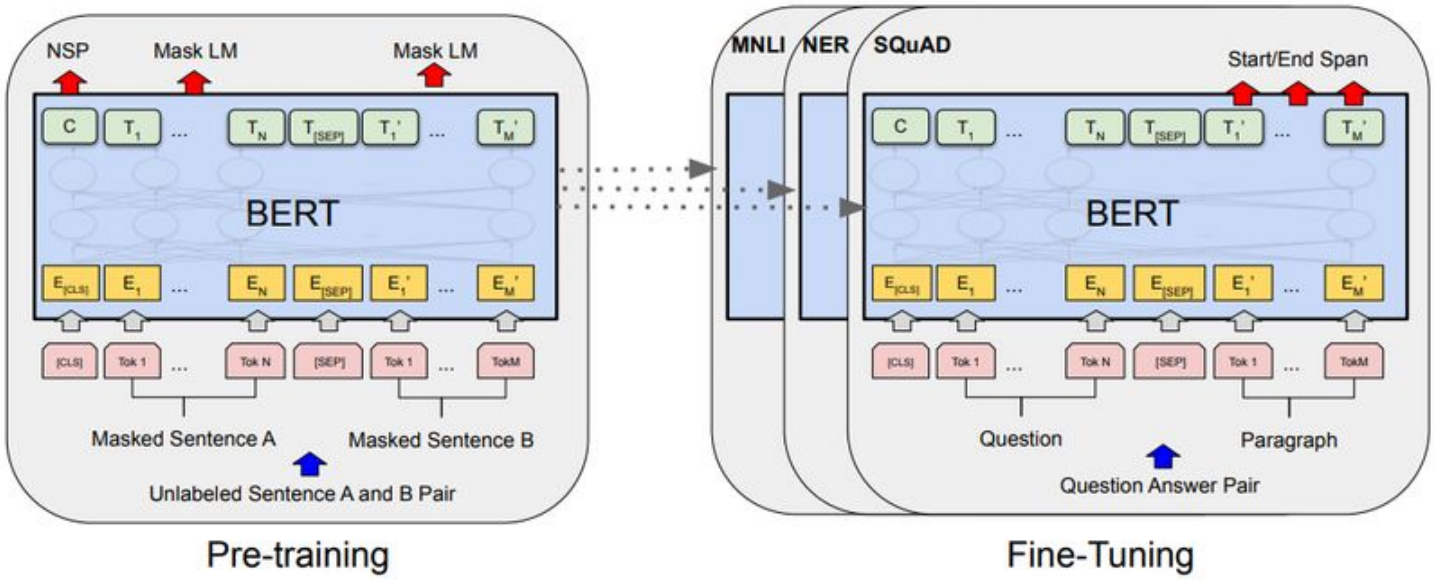
SHAP feature importance can be explained using a waterfall plot



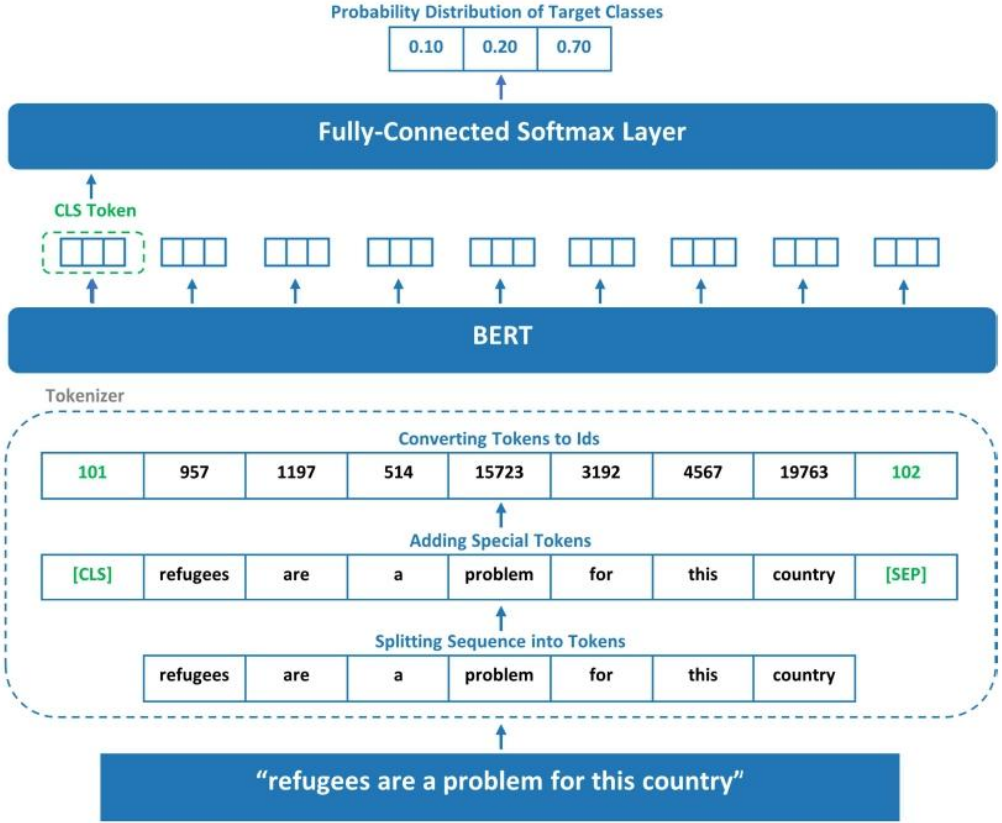
High-level visualization of the transformer architecture



High-level visualization of the BERT architecture for Question Answering



High-level visualization of the model architecture

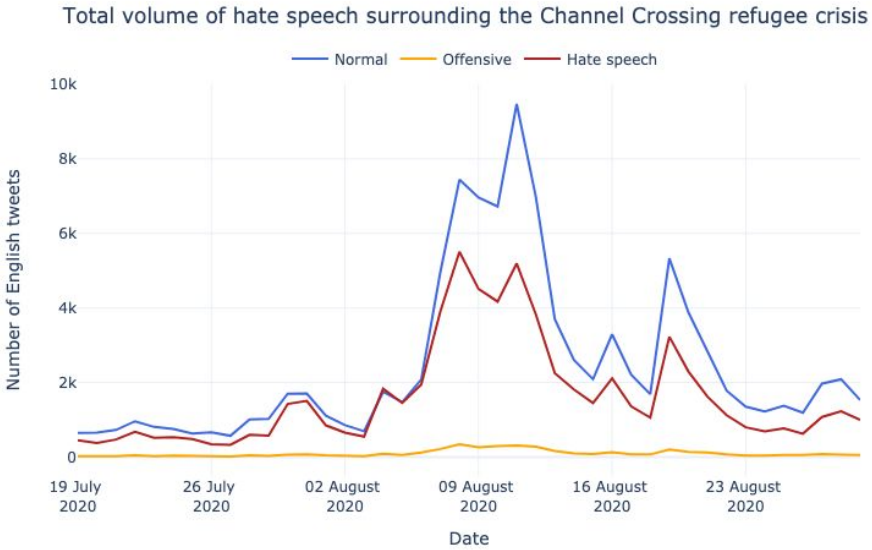


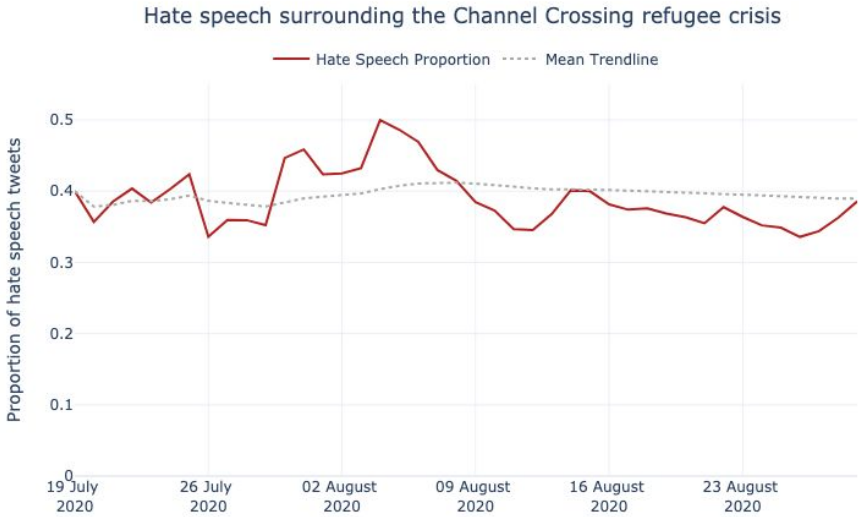
Each crisis have very different levels of hate speech

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

Hate Speech Proportion

United Kingdom Channel Crossings

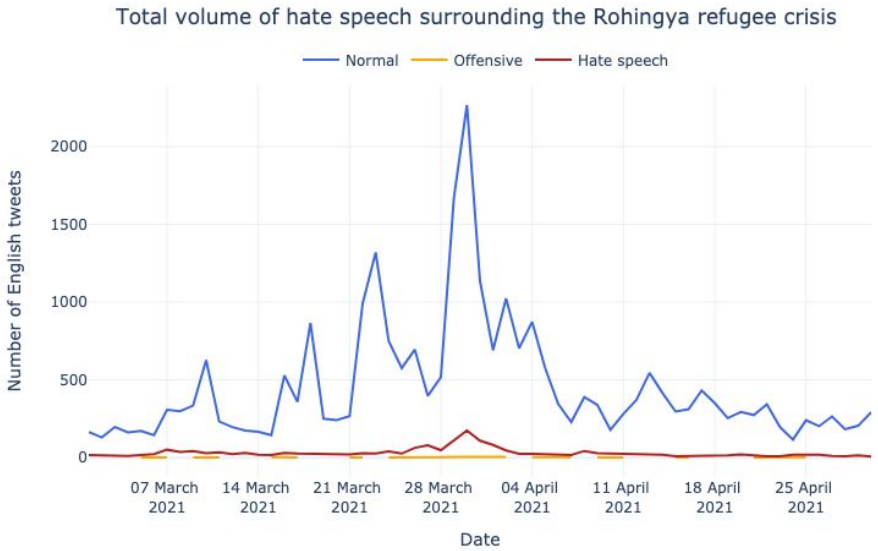


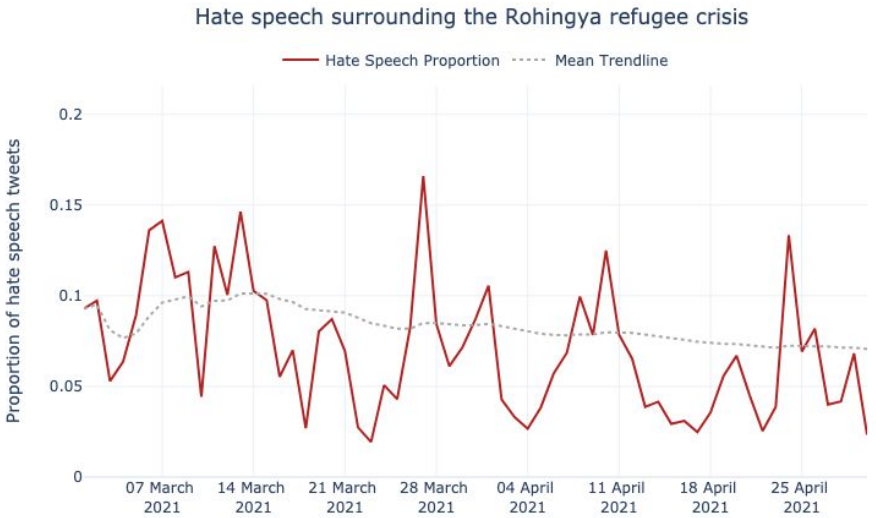


United Kingdom Channel Crossings

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

Rohingya

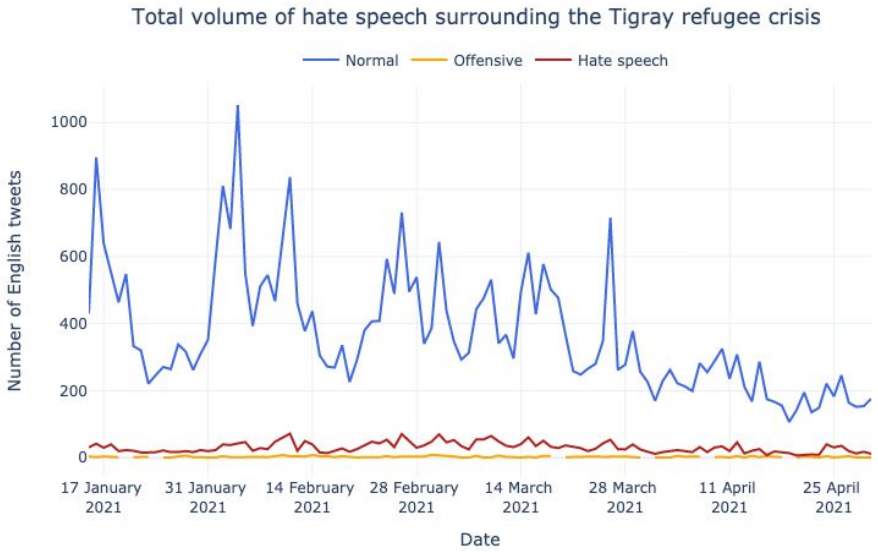

Total volume of hate speech surrounding the Rohingya refugee crisis
Normal
Offensive
Hate speech
Number of English tweets
2000
1500
1000
500
0
07 March 2021
14 March 2021
21 March 2021
28 March 2021
04 April 2021
11 April 2021
18 April 2021
25 April 2021
Date

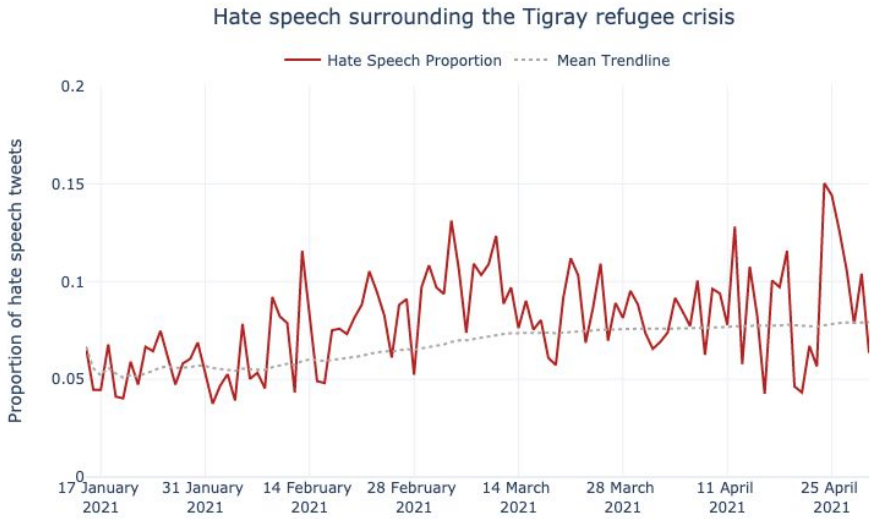

Hate speech surrounding the Rohingya refugee crisis
Hate Speech Proportion
Mean Trendline
Proportion of hate speech tweets
0.2
0.15
0.1
0.05
0
07 March 2021
14 March 2021
21 March 2021
28 March 2021
04 April 2021
11 April 2021
18 April 2021
25 April 2021

Rohingya Channel Crossings

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

Tigray


Total volume of hate speech surrounding the Tigray refugee crisis
Normal
Offensive
Hate speech
Number of English tweets
1000
800
600
400
200
0
17 January 2021
31 January 2021
14 February 2021
28 February 2021
14 March 2021
28 March 2021
11 April 2021
25 April 2021
Date


Hate speech surrounding the Tigray refugee crisis
Hate Speech Proportion
Mean Trendline
Proportion of hate speech tweets
0.2
0.15
0.1
0.05
0
17 January 2021
31 January 2021
14 February 2021
28 February 2021
14 March 2021
28 March 2021
11 April 2021
25 April 2021

Tigray Channel Crossings

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