Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Language Technology

DiGiTAL TOTAL

Multilingual Racial Hate Speech Detection Using Transfer Learning

Abinew Ali Ayele[1,2], Skadi Dinter[1], Seid Muhie Yimam[1], Chris Biemann[1]

[1]Language Technology Group, Department of Informatics, Universität Hamburg, Hamburg, Germany
[2]Faculty of Computing, BiT, Bahir Dar University, Bahir Dar, Ethiopia

{abinew.ali.ayele, seid.muhie.yimam,chris.biemann}@uni-hamburg.de, skadi.dinter@posteo.de

Abstract

- Social media eases the spread of hateful or racist content.
- We employ **Yandex Toloka** crowdsourcing annotation platform.
- Annotate **5k** tweets into **hate**, **offensive** or **normal** categories and further identify offensive & hateful tweets as **racial** or **non-racial**.
- We apply transfer learning by fine-tuning the **HateXplain** model based on **multilingual BERT** and **CamemBERT**.
- CamemBERT yields the best results and able to resolve annotation ties in our experiments.

Introduction

- No common definition for hate speech
- **Hate Speech:** hatred expressions targeting group identities such as race, color, sexual orientation, religion, etc.…

Research Questions

- Can BERT and HateXplain models be efficiently adapted to other languages or cultures, specifically to racial hate speech detection tasks in French?
- What are the main challenges of racial hate speech data annotation on Toloka crowdsourcing platform?

Main Contributions

- Collections of **French racial** hate speech lexicon entries and dataset.
- Exploring the annotation challenges of racial hate speech on the Yandex Toloka crowdsourcing platform.
- Adaptation of a racial hate speech detection model for the French Twitter dataset.

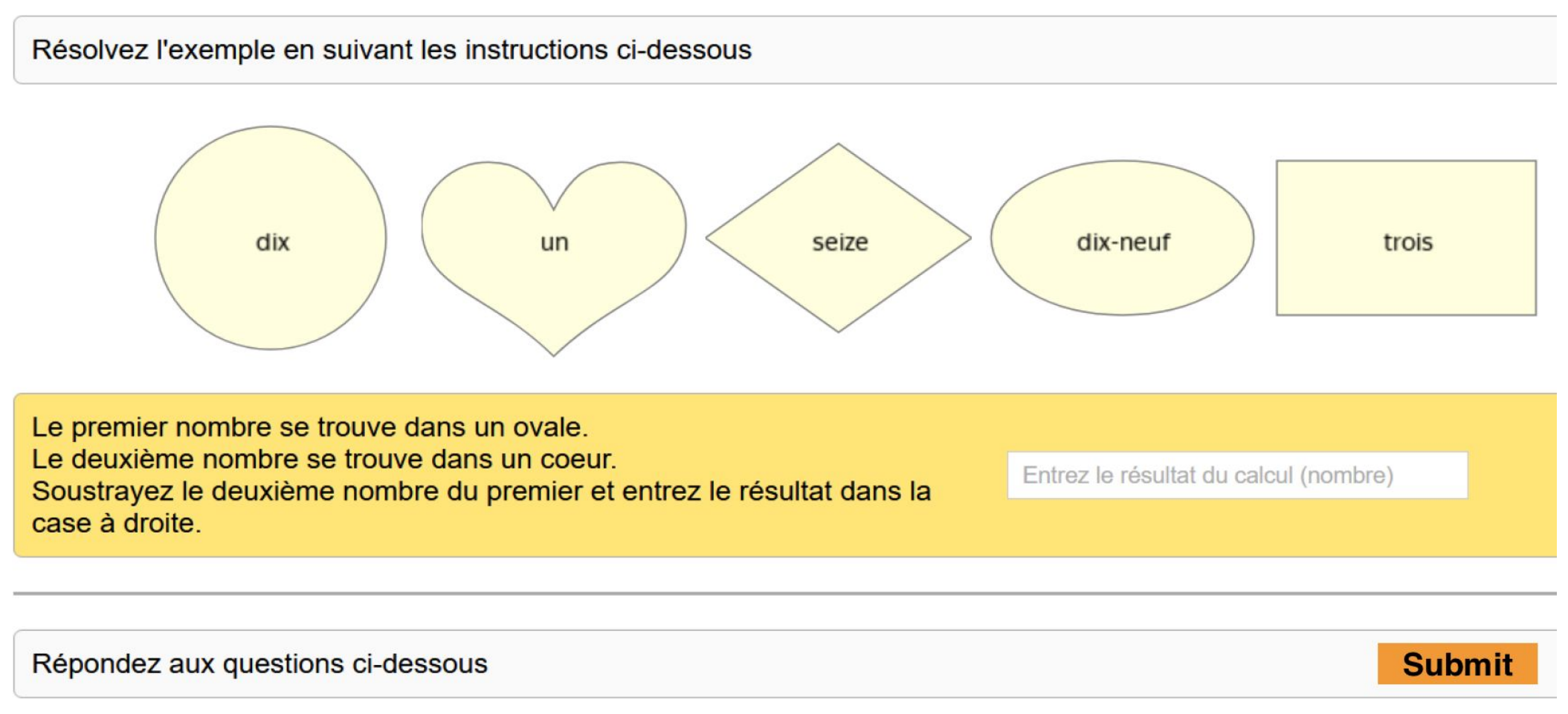
Data Collection

- Source: Tweets **May 25 – June 25, 2020**, after the death of **G. Floyd.**
- Collected **3,473 French hate speech lexicon** entries.
- Apply Pycld2 to filter French tweets.
- Truncated tweets are removed
- Usernames and URLs are anonymized as <USER> and <URL>.

Annotation

- Annotation Tool: Toloka
- 5k tweets annotated by **3 independent performers**
- Gold label: determined with **majority voting**
- Labels:
  - Hate, Offensive, Normal, Unsure
  - Hate & Offensive {Racial, Non-racial}
- Fleiss Kappa: **0.34**
- Each annotator earned **$0.1 per task**
- Control Questions:
  - 50 random tweets are annotated and evaluated by experts for correctness.
  - Each Toloka task contained **15 tweets,**1 of them is a **control question** to control malicious performers.
- Before joining the main task, performers are given:
  - Annotation guidelines
  - Two training task pools to be completed successfully.
  - A French language test as presented below.

Sample French Language Test


Résolvez l'exemple en suivant les instructions ci-dessous

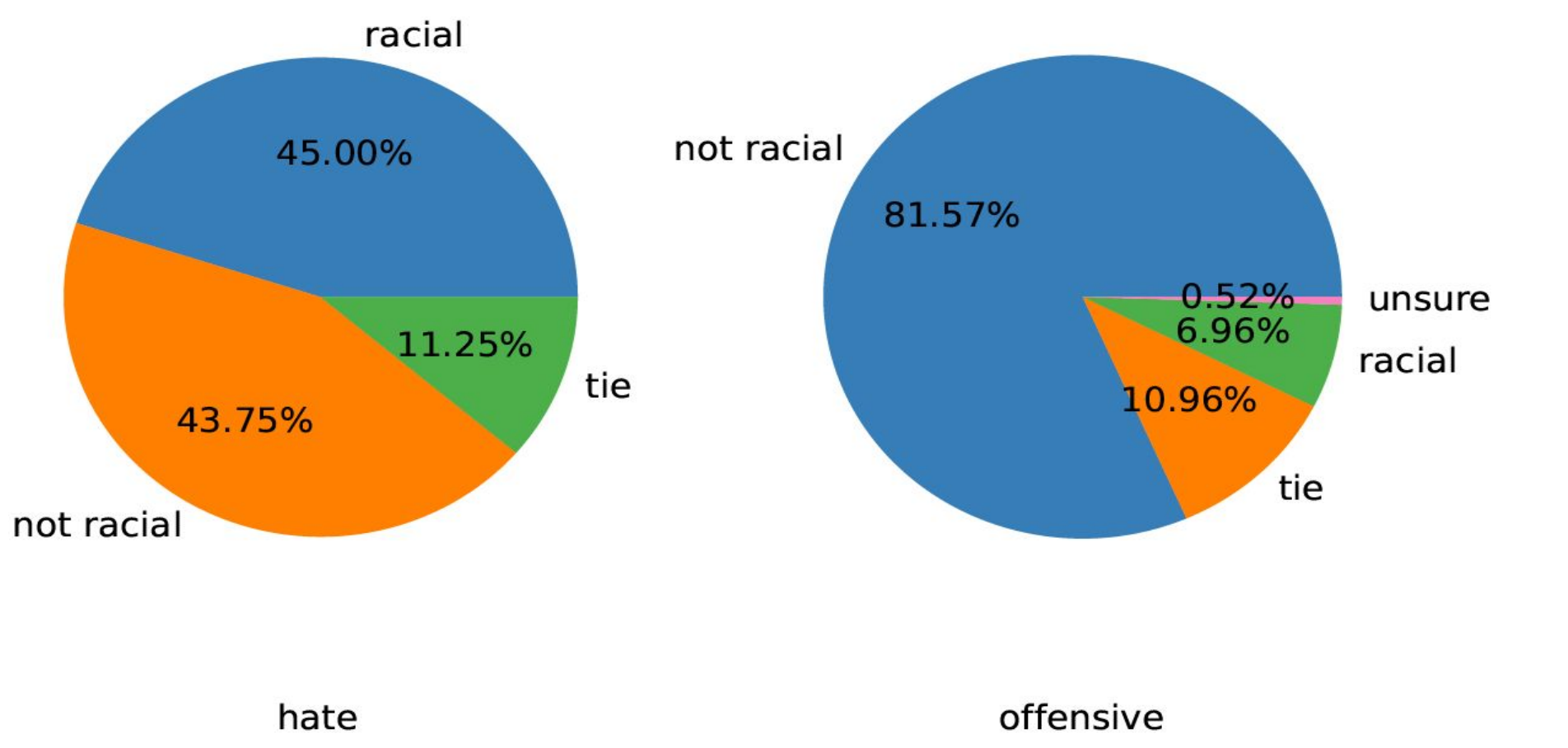
Le premier nombre se trouve dans un ovale.
Le deuxième nombre se trouve dans un coeur.
Soustrayez le deuxième nombre du premier et entrez le résultat dans la case à droite.

Entrez le résultat du calcul (nombre)

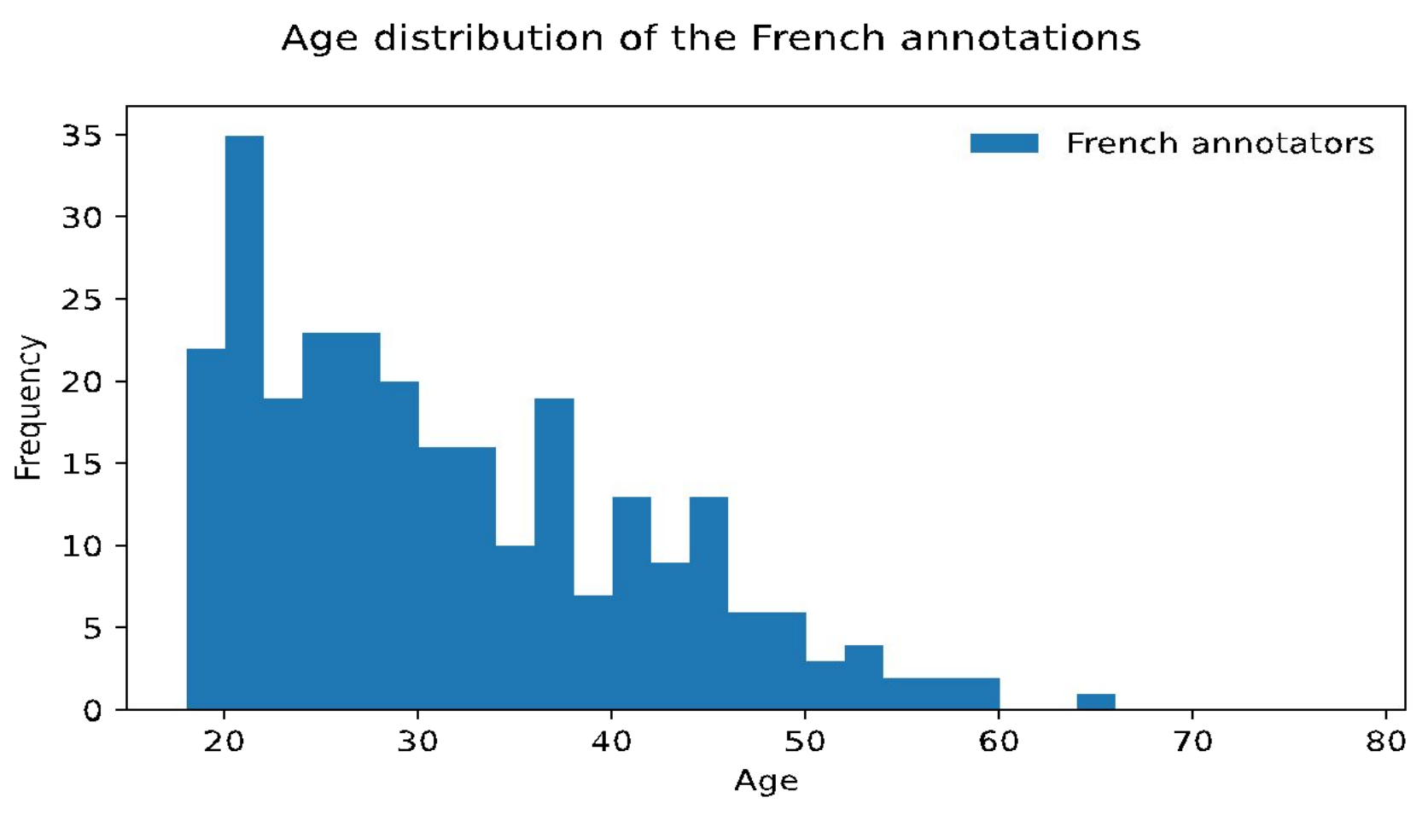
Répondez aux questions ci-dessous

Submit

Label Distribution in the Dataset



Age Distribution of Performers

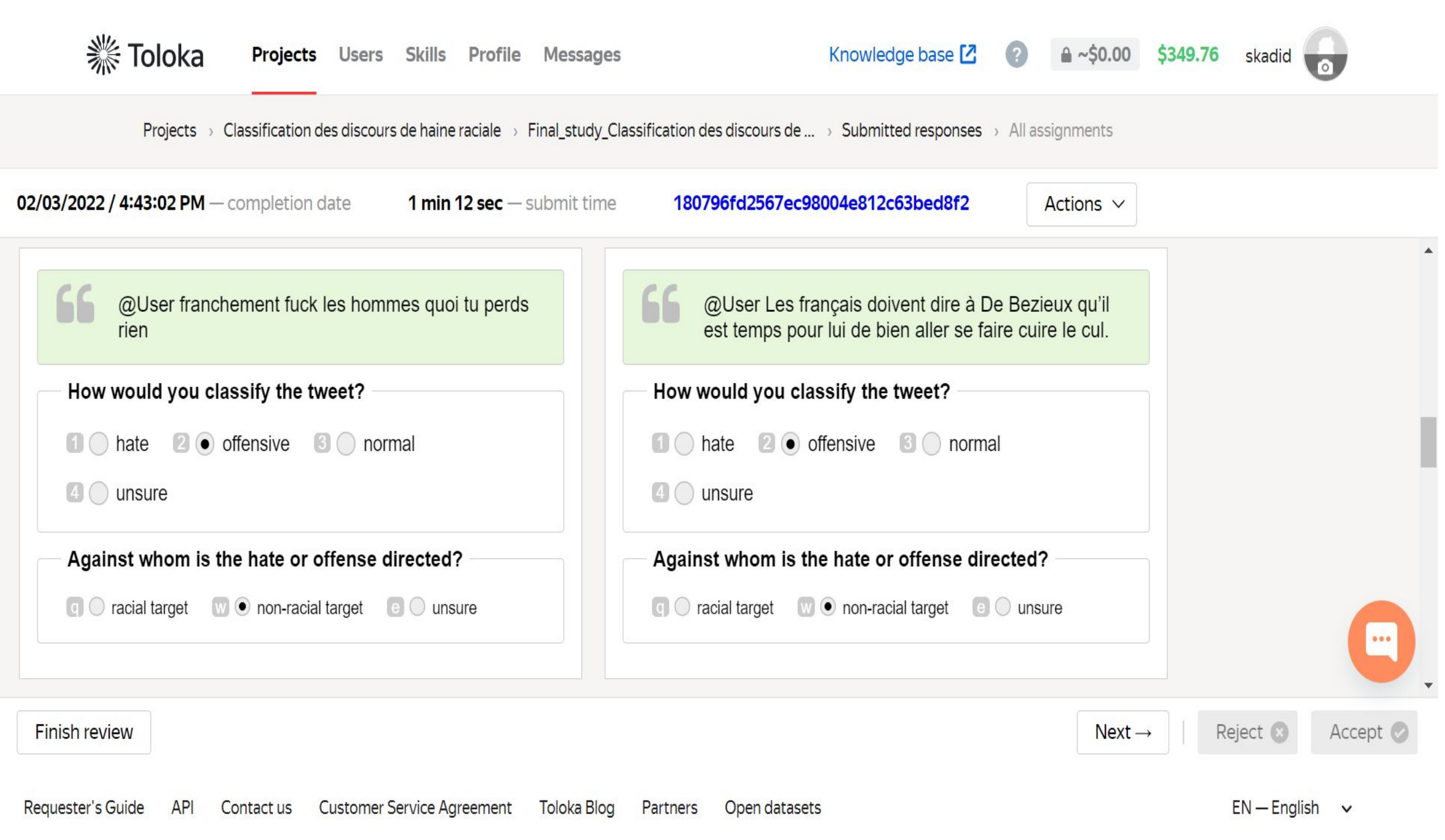


Overall Annotation Information

Fleiss Kappa score	0.3
Total number of Annotated tweets	5002
Number of annotators participated in the task	275
Mean age of annotators in years	31.11
Country distribution of annotators	265 Fr, 8 Be, 3 O
Accuracy for 50 random tweets	0.24
F1 score for 50 random tweets	0.24
Racial accuracy for 50 random tweets	0.12
Average time for 15 tweets	2 min 10 sec
Number of collected keywords	3473

Keys: **Fr**= French, Be= Belgium, O = Others

Sample Task Presented to Toloka Performers



Projects › Classification des discours de haine raciale › Final_study_Classification des discours de ... › Submitted responses › All assignments

02/03/2022 / 4:43:02 PM — completion date 1 min 12 sec — submit time 180796fd2567ec98004e812c63bed8f2

@User franchement fuck les hommes quoi tu perds rien

How would you classify the tweet?
1 ○ hate 2 ● offensive 3 ○ normal
4 ○ unsure

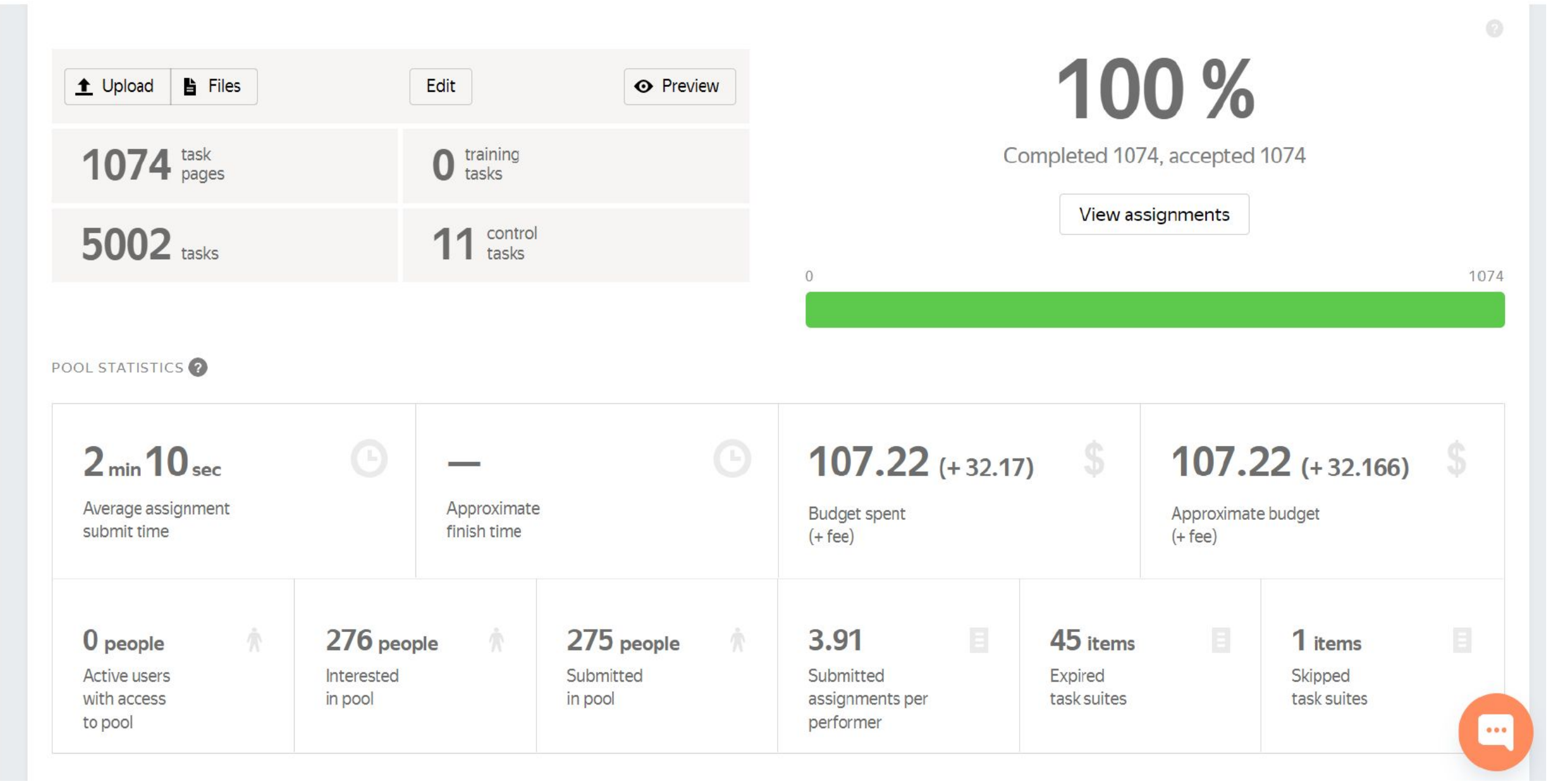
Against whom is the hate or offense directed?
q ○ racial target w ● non-racial target e ○ unsure

@User Les français doivent dire à De Bezieux qu'il est temps pour lui de bien aller se faire cuire le cul.

How would you classify the tweet?
1 ○ hate 2 ● offensive 3 ○ normal
4 ○ unsure

Against whom is the hate or offense directed?
q ○ racial target w ● non-racial target e ○ unsure

Completed Annotation Project on Toloka



1074 task pages
0 training tasks
5002 tasks
11 control tasks

100 %
Completed 1074, accepted 1074

POOL STATISTICS

2 min 10 sec Average assignment submit time
— Approximate finish time
107.22 (+ 32.17) Budget spent (+ fee)
107.22 (+ 32.166) Approximate budget (+ fee)
0 people Active users with access to pool
276 people Interested in pool
275 people Submitted in pool
3.91 Submitted assignments per performer
45 items Expired task suites
1 items Skipped task suites

Experimental Results

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

Further Experiments Based on Exp. 2.1 above

Experiment	Accuracy	F1	Epochs	Learn. rate
2.1 a)	0.886	0.859	3	5e-5
2.1 b)	0.899	0.882	2	5e-5
2.1 c)	0.888	0.876	1	5e-5
2.1 d)	0.882	0.869	4	5e-5
2.1 e)	0.852	0.784	3	5e-4
2.1 f)	**0.892**	**0.869**	**3**	**5e-6**
2.1 g)	**0.892**	**0.874**	**4**	**5e-6**

Conclusion and Future Works

- BERT model is successfully fine-tuned with the dataset, and with the translated HateXplain dataset.
- We achieved **88% accuracy** & **86% F1-score,** and are improving over the baseline HateXplain model.
- In future:
  - Improve **data selection strategies** to reduce the **class imbalance** problem.
  - Explore **targets** and label decision **rationales**