IMSyPP
Innovative Monitoring Systems and
Prevention Policies of Online Hate Speech

WP2 – Hate speech detection and trends

Multilingual Hate Speech Database

February, 2021

FUNDING:	European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263)
TYPE OF ACTION:	REC-AG Rights, Equality and Citizenship Programme Action Grant Directorate-General Justice and Consumers
CALL:	REC-RRAC-RACI-AG-2019 Call for proposals to prevent and combat racism, xenophobia, homophobia and other forms of intolerance and to monitor, prevent and counter hate speech online
PRIORITY:	To monitor, prevent and counter hate speech online
DURATION:	March 1, 2020 – February 28, 2022

AUTHORS:
Petra Kralj Novak, Jožef Stefan Institute
Igor Mozetič, Jožef Stefan Institute
Guy De Pauw, TextGain BBVD
Matteo Cinelli, Ca'Foscari University of Venice

Jožef Stefan
Institute
Ljubljana, Slovenia

Executive summary

In this document we present the IMSyPP Multilingual Hate Speech Database. We describe the collection, selection, annotation and agreement of the social media data collected. Up to our knowledge, it is the only dataset with context information. The Twitter part of the dataset is published and available on the clarin.si language resources repository:

https://www.clarin.si/repository/xmlui/handle/11356/1398

Table of contents

1 Relevance for IMSyPP

The work presented in this deliverable is an essential part of the IMSyPP project. It addresses directly the Need 5 in the Description of Action: Resources for science, and is a necessary step towards the objective of Tracking of hate speech trends online. It is the result of Task 2.1 Data acquisition and Task 2.2 Annotation of online comments. The tasks were led by JSI with collaboration from UNIVE and TEXTGAIN BVBA. The results are a direct input for Task 2.3: Hate speech detection modelling, and consequently for most of the other tasks in the project.

The goal of this deliverable was to develop high quality, large datasets of examples of hate speech. The datasets are used to train machine learning algorithms that are in turn used to perform hate speech classification. Based on the results presented, the goal was successfully achieved.

2 Datasets

Four annotated datasets were created, one for each target language: English, Italian, Slovenian and Dutch. Each dataset is unique in some aspects. We also promptly reacted to the emergence of the Covid-19 pandemic and focused some of our data collection efforts towards this topic.

The Italian and English datasets are unique as they originate from YouTube, a social media platform which is not commonly analyzed for hate speech detection. Additionally, both datasets include contextual information in the form of annotated threads of YouTube comments which is not available on other social media, e.g., Twitter.

The Slovenian dataset consists of Twitter posts and was drawn from an exhaustive set of all Slovenian Twitter posts of the last three years. The dataset is not focused on any specific topic, but reflects the increased engagement of Twitter users during the emergence of the Covid-19 pandemic.

The Dutch dataset consists of Twitter posts, Facebook and YouTube comments centered around a series of thematic and regional clusters. Additional data was collected from the popular forums GeenStijl and Dumpert. All data sources are threaded with the exception of the Twitter data.

The Slovenian Twitter annotated dataset is publicly available in the CLARIN repository at: https://www.clarin.si/repository/xmlui/handle/11356/139

The other datasets are not publicly available due to the GDPR and Terms of service restrictions. They can be made available for research purposes on the basis of individual agreements.

2.1 Dataset properties

Tables 1 and 2 summarize the properties of the four datasets.

Language	**Source**	**Topic**	**Dates**
English	YouTube comments	Covid-19	Feb. 2020 - May 2020
Italian	YouTube comments	Covid-19	Jan. 2020 - May 2020
Slovenian	Twitter posts	General	Dec. 2017 - Oct. 2020
Dutch	Twitter, Facebook, YouTube, GeenStijl, Dumpert	General, Covid-19	Jan 2018 - Oct 2020

Table 1: Datasets' properties in terms of data sources, topics covered and timeframe.

	Size	Number of annotators	Inter-annotator agreement on hate speech type (Krippendorff Alpha)	Inter-annotator agreement on hate speech target (Krippendorff Alpha)
English				
Train	103,190	10	0.591	0.463
Evaluation	10,759	10	Annotation in progress	Annotation in progress
Italian				
Train	119,670	8	0.586	0.617
Evaluation	21,072	8	0.555	0.367
Slovenian				
Train	99,809	10	0.606	0.645
Evaluation	20,000	10	0.536	0.503
Dutch				
Train	26,031	8	Annotation in progress	Annotation in progress
Evaluation	3,000	8	(Kappa) 68.6	(Kappa) 65.2

Table 2: Annotation properties in terms of dataset sizes and inter-annotator agreements.

2.2 Annotation procedure

The annotation procedure consists of selecting the data, setting up the annotation platform, recruiting and training the annotators, monitoring the annotators' progress and agreement, and resolving severe disagreement between the annotations.

For each dataset, a separate set of data was selected and annotated for training and evaluating machine learning models. The training data selection was optimized to get hate speech rich (biased) training datasets to be used by machine learning algorithms. The evaluation set data selection targeted a random sample of the data to be used to evaluate the performance of the trained model on real data.

We developed a simple but effective annotation platform in Google Sheets with drop-down menus for quick annotation (See the Annotation guidelines in the Appendix). Google Sheets allows for programming access which was used to upload the data, set up the interface and to download the annotated data. On the user side, it is customizable by the user (font size, column width) and allows to use browser plugins like Read Aloud to help the annotators with the reading.

Annotators were recruited in Slovenia, Italy and Belgium. Good knowledge of the target language (native speakers of Slovenian, Dutch and Italian and proficient users of English) as well as expressed interest in the hate speech domain were required. The annotators were mostly PhD and Master's students of social sciences. Annotators were provided with written annotations guidelines in their mother tongue (See the Appendix), and a videoconference lecture with oral instructions and a demonstration of the annotation platform.

The annotators were working remotely on their own schedule. Rough deadlines were set to discourage procrastination. The progress in terms of the number of annotations and agreement between annotators was monitored regularly. We monitored the following:

- Number of annotations
- Inter-agreement accuracy and matrix on hate speech type
- Self-agreement accuracy and matrix on hate speech type
- Inter and self-agreement on hate speech target
- Nominal, Interval and Ordinal Krippendorff Alpha on hate speech type

For some datasets, once annotators completed their task, a special session was held to resolve the cases of severe disagreements.

3 Dataset specific details

3.1 English YouTube comments

We annotated English YouTube comments for hate speech type and hate speech target. Two sets were annotated: a training set with 51,665 comments and an evaluation set with 10,759 comments (in progress). The comments to be annotated were sampled from the English YouTube comments on videos about the Covid-19 pandemic. The comments and the videos metadata were collected using the YouTube API.

3.1.1 Training dataset

16,904 videos with 5,503,283 comments were collected in the period from February 2020 to April 2020. The distribution of the number of comments per video is presented in Figure 1.

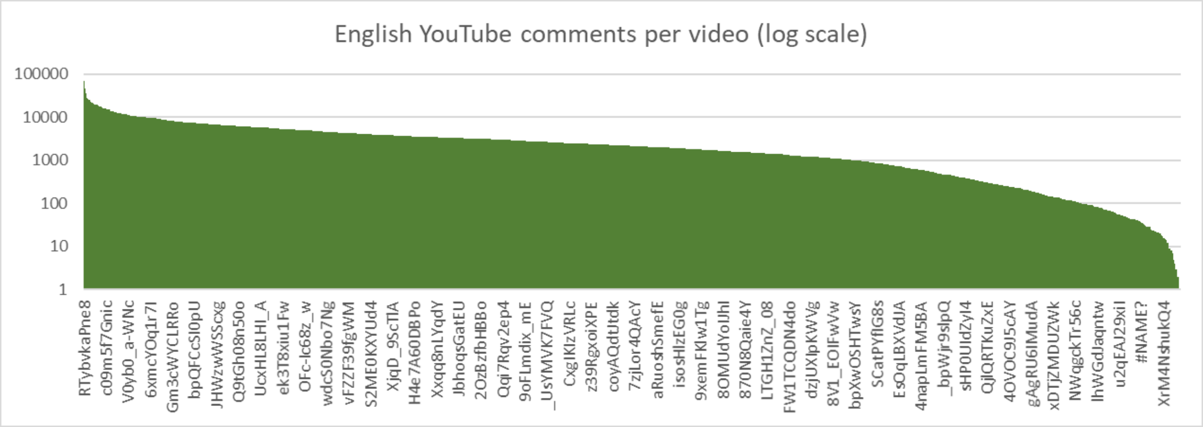


Figure 1: Number of comments per video in our English YouTube collected sample on Covid-19 (logarithmic scale).

In order to get a training set that is rich with hate speech, we implemented a preprocessing step consisting in the annotation of the whole set of comments by means of a (basic) hate speech classifier (machine learning model) that assigns a score between -3 (hateful) and +3 (normal) trained on FRENK English data[1]. Even though the basic model is not very accurate, its performance is better than random and we used its result for selecting the training data to be annotated and later used for training machine learning models.

[1] Ljubešić, N., Fišer, D., & Erjavec, T. (2019, September). The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English. In International Conference on Text, Speech, and Dialogue (pp. 103-114). Springer, Cham.

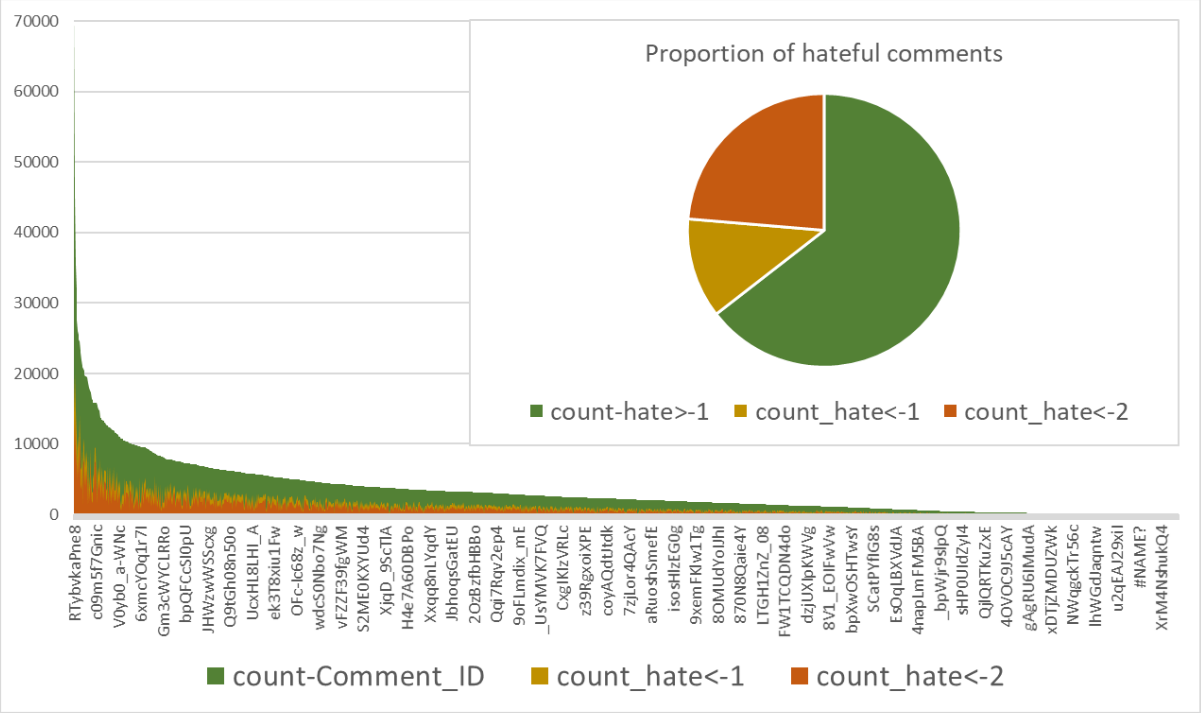


Figure 2: Hate speech scores by out FRENK model per threads length.

Sampling. We selected the videos that have between 10 and 2000 comments and the percent of hateful comments with a score below -3 of at least 30%. This resulted in 74 videos with 51,665 comments in total.

Dividing the comments (threads) between the annotators:

- There are 10 annotators
- Each comment should be annotated twice by two different annotators
- Each annotator should get approximately the same number of comments to annotate
- Each pair of annotators should have approximately the same overlap
- Each annotator should have both long and short threads

The overlap between the annotators is shown in Table 3, and the distribution of thread lengths are in Figure 3.

	1	2	3	4	5	6	7	8	9	10	Grand To
1		1961	1135	1114	465	462					5137
2			1931	1143	1063	733	451				5321
3				1783	1154	993	725	406			5061
4					1828	1207	976	614	510	234	5369
5	188					1693	1220	920	614	718	5353
6	283					224	1613	1376	910	628	5034
7	1037	227			42			1678	1393	868	5245
8	1049	856	232						1583	1444	5164
9	1496	1046	842		325					1570	5279
10	819	839	1507	1527							4692
Grand To	4872	4929	5647	5567	4877	5312	4985	4994	5010	5462	51655

Table 3: Overlap between annotators in the English YouTube training dataset (each comment is counted once).

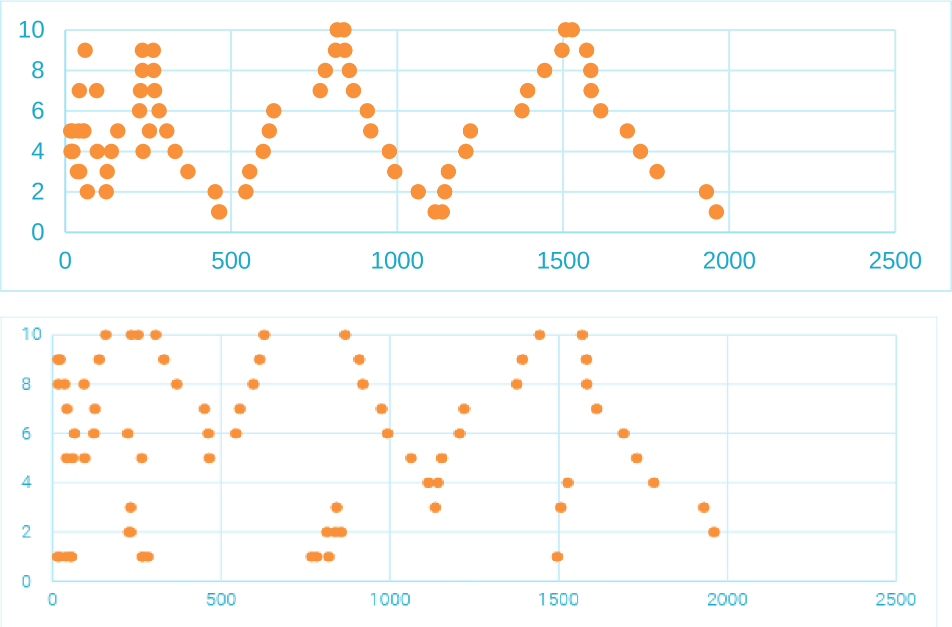


Figure 3: Distribution of thread lengths (X axis) per annotator (Y axis) for the unique comments (top) and for the replicas (bottom).

3.1.2 Evaluation dataset

Evaluation dataset should be disjoint from the training dataset to ensure proper evaluation. 3,144 videos with 2,052,784 comments were collected in the first week of May 2020. The distribution of the number of comments per video is presented in Figure 4.

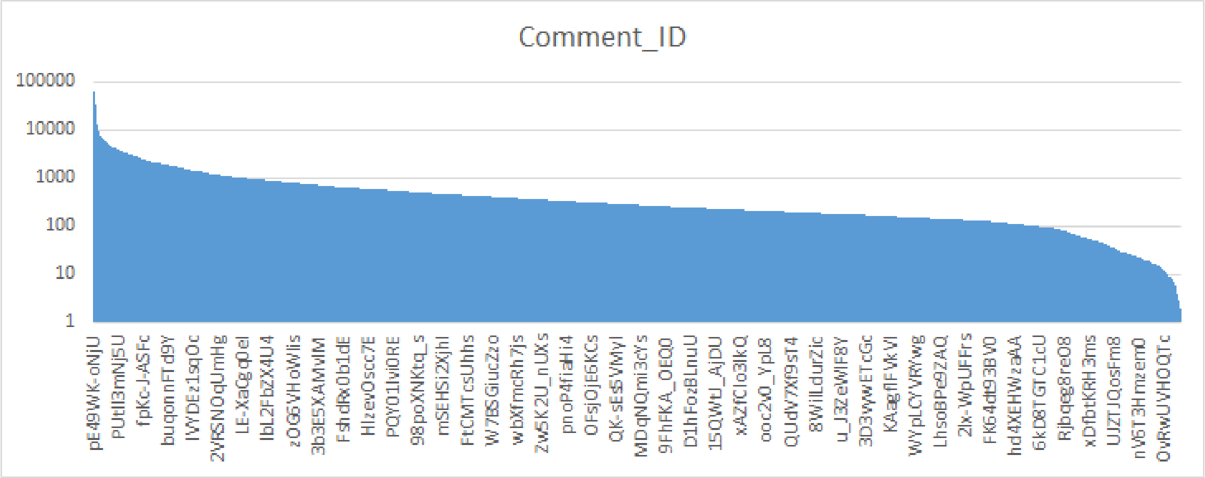
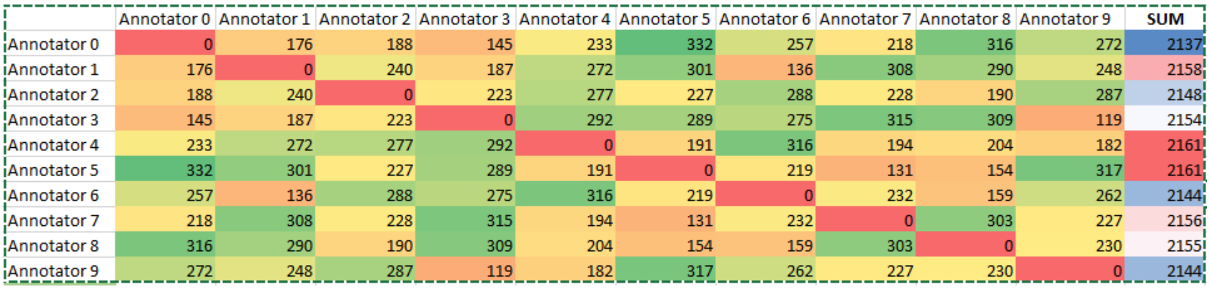


Figure 4: Distribution of thread lengths in the collected YouTube English dataset to be sampled for the evaluation set (logarithmic scale).

We have sampled 100 posts of lengths varying between 10 and 200 to achieve an evaluation set size of 10,759 YouTube comments. Following the same criteria as for the training set each annotator got about 2,150 comments to annotate. The distribution of the number of comments between pairs of annotators is presented in Table 4.



	Annotator 0	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Annotator 6	Annotator 7	Annotator 8	Annotator 9	SUM
Annotator 0	0	176	188	145	233	332	257	218	316	272	2137
Annotator 1	176	0	240	187	272	301	136	308	290	248	2158
Annotator 2	188	240	0	223	277	227	288	228	190	287	2148
Annotator 3	145	187	223	0	292	289	275	315	309	119	2154
Annotator 4	233	272	277	292	0	191	316	194	204	182	2161
Annotator 5	332	301	227	289	191	0	219	131	154	317	2161
Annotator 6	257	136	288	275	316	219	0	232	159	262	2144
Annotator 7	218	308	228	315	194	131	232	0	303	227	2156
Annotator 8	316	290	190	309	204	154	159	303	0	230	2155
Annotator 9	272	248	287	119	182	317	262	227	230	0	2144

Table 4: Overlap in number of comments between pairs of annotators (each comment is counted twice).

The thread lengths per annotator are presented in Figure 5.

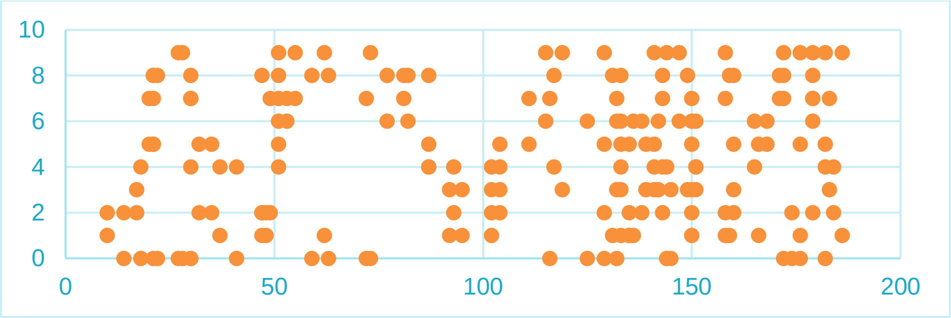


Figure 5: Thread length (X axis) per annotator (Y axis) in the English evaluation dataset.

3.1.3 Annotation results

The overall numbers of the different types of hate speech in our English dataset are in Table 5.

Annotated for Type:	**103190**
0. appropriate	52990
1. inappropriate	1739
2. offensive	45863
3. violent	2589
Annotated for Target:	**48291**
1. racism	3656
2. migrants	81
3. islamophobia	1438
4. antisemitism	24
5. religion	309
6. homophobia	8
7. sexism	92
8. ideology	1023
9. media	4907
10. politics	20754
11. individual	10865
12. other	5134

Table 5: The numbers of different types and targets of hate speech in the English training set.

3.2 Italian YouTube comments

We collected and annotated a large set of Italian YouTube comments for hate speech type and hate speech target. The comments to be annotated were sampled from the Italian YouTube comments on videos about the Covid-19 pandemic in the period from January 2020 to May 2020. The comments and the videos metadata were collected using the YouTube API.

Two sets were annotated: a training set with 59,870 comments and an evaluation set with 10,536 comments.

3.2.1 Training dataset

- All videos: 26.267
- All comments: 1.273.930

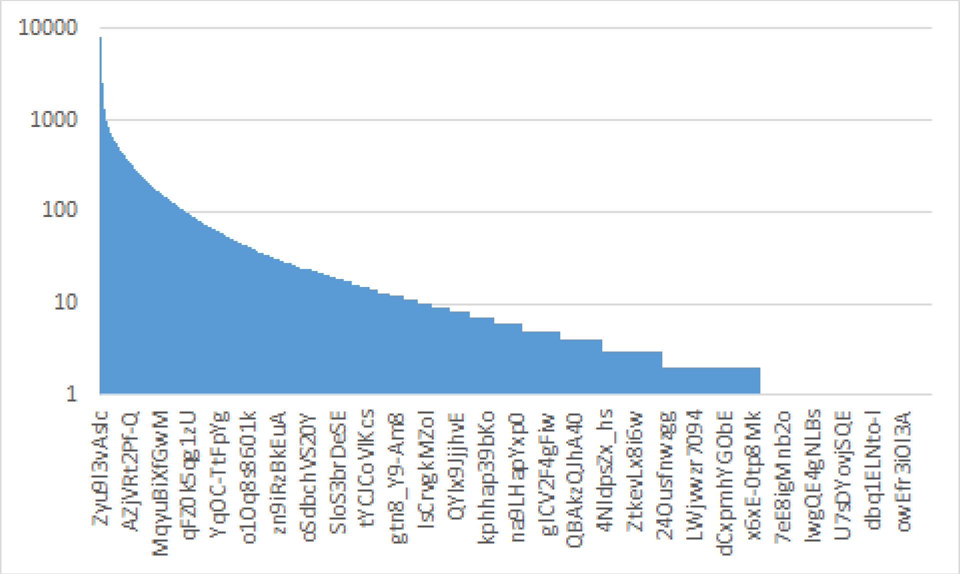


Figure 3: Distribution of the Italian comments per YouTube video (logarithmic scale) in the training dataset.

In order to get a training set that is rich with hate speech, we annotated all the comments with a (basic) hate speech classifier (machine learning model) that assigns a score between -3 (hateful) and +3 (normal). The basic classifier was trained on publicly available dataset of Italian hate speech against migrants. Even though the basic model is not very accurate, its performance is better than random and we used its result for selecting the training data to be annotated and later used for training machine learning models.

Sampling. The threads (with comments) were selected according to the following criteria:

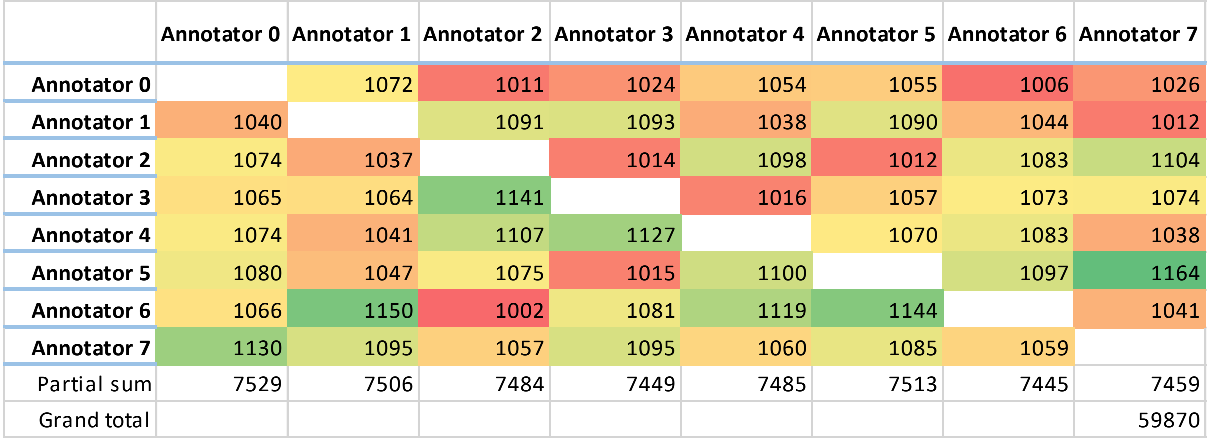
- No. of comments in a thread >= 10
- No. of comments in a thread < 500
- Probability of hate-2 > 0.05

The application of these criteria resulted in 1.168 threads (VideoIds) and 59.870 comments. In this selection, there are 13.749 comments with hate speech score below -1 and 4.378 comments with the score below -2. In all the selected threads, the difference between the score of the most positive and most negative comment is about 4.5.

Criteria for dividing the comments (threads) between the annotators:

- There are 8 annotators
- Each comment should be annotated twice by two different annotators
- Each annotator should get approximately the same number of comments to annotate
- Each pair of annotators should have approximately the same overlap
- The threads should remain intact
- Each annotator should have both long and short threads

The results of the distribution of the comments between the annotators by the above criteria are in Figure 6, and the thread lengths are in Figure 7.



	Annotator 0	Annotator 1	Annotator 2	Annotator 3	Annotator 4	Annotator 5	Annotator 6	Annotator 7
Annotator 0		1072	1011	1024	1054	1055	1006	1026
Annotator 1	1040		1091	1093	1038	1090	1044	1012
Annotator 2	1074	1037		1014	1098	1012	1083	1104
Annotator 3	1065	1064	1141		1016	1057	1073	1074
Annotator 4	1074	1041	1107	1127		1070	1083	1038
Annotator 5	1080	1047	1075	1015	1100		1097	1164
Annotator 6	1066	1150	1002	1081	1119	1144		1041
Annotator 7	1130	1095	1057	1095	1060	1085	1059	
Partial sum	7529	7506	7484	7449	7485	7513	7445	7459
Grand total								59870

Figure 4: Overlap between annotators in the Italian YouTube training dataset (each comment is counted once).

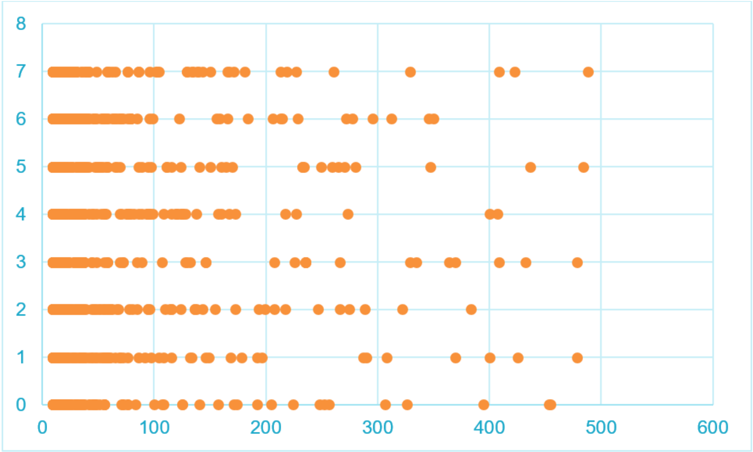


Figure 5: Thread lengths per annotator in the Italian YouTube comments training dataset.

3.2.2 Evaluation dataset

The evaluation set for Italian was collected analogously to the English evaluation dataset. Data was collected in May 2020 and a random (unbiased) sample of 10,543 comments grouped into 151 threads (videos) was split among eight annotators. Each comment was annotated twice by two different annotators. The splitting procedure was optimized to get approximately equal overlap (in the number of comments) between each pair of annotators.

The annotation procedure resulted in 21,072 annotations for Type (Tipo) and 3,929 annotations for Target.

3.2.3 Annotation results

The annotation results for the Italian training and evaluation sets are summarized in Table 6.

Training set		**Evaluation set**	
Annotated for Tipo:	119670	**Annotated for Tipo:**	21072
0. appropriato	77718	0. appropriato	15956
1. inappropriato	5447	1. inappropriato	770
2. offensivo	32712	2. offensivo	4082
3. violento	3793	3. violento	264
Annotated for Target:	32859	**Annotated for Target:**	3929
1. razzismo	2080	1. razzismo	122
2. migranti	886	2. migranti	10
3. islamofobia	41	3. islamofobia	0
4. antisemitismo	24	4. antisemitismo	10
5. religione	120	5. religione	24
6. omofobia	25	6. omofobia	1
7. sessismo	232	7. sessismo	49
8. ideologia	2303	8. ideologia	96
9. media	1229	9. media	349
10. politica	15476	10. politica	1656
11. individuo	5409	11. individuo	823
12. altro	4576	12. altro	751
13. nord vs. sud	458	13. nord vs. sud	38

Table 6: The numbers of different types (tipo) and targets of hate speech in the Italian training and evaluation datasets.

3.3 Slovenian Twitter posts

We collected almost three years of all Slovenian Twitter data in the period from December 1, 2017 to October 1, 2020, in total 11,135,654 tweets. The period includes several government changes, elections and the first Covid-19-related lockdown.

The Twitter data was collected by the TweetCat tool[2]. The TweetCat tool is focused on harvesting Twitter data of less frequent languages by continuously searching for new users tweeting in the language of interest by querying the Twitter Search API for the most frequent and unique words in that language. Once a series of new potential users tweeting in the language of interest are identified, their full timeline is retrieved and language identification is run over their timeline. If a specific user shows to tweet predominantly in the language of interest, they are added to the user index and their tweets are collected for the remainder of the collection procedure. In our case, the collection procedure has started end of 2017 and is still running. Given that we are building the Slovene Twitter user index with a previous version of the tool since 2013, we are very confident that we have the full Slovene Twittosphere covered.

3.3.1 Training dataset

The training set is sampled from data collected before February 2020. The sampling was intentionally biased to contain as much hate speech as possible. A simple model was used to flag potential hate speech content and additionally, filtering by users and by tweet length (number of characters) was applied. About 50,000 tweets were selected.

3.3.2 Evaluation dataset

The evaluation set is sampled from data collected between February 2020 and August 2020. Contrary to the training set, the evaluation set is an unbiased random sample. Since the evaluation set is from a

[2] N. Ljubešić, D. Fišer, T. Erjavec, TweetCaT: a tool for building Twitter corpora of smaller languages, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 2279–2283. URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/834_Paper.pdf

later period compared to the training set, the possibility of data linkage is minimized. Furthermore, the estimates of model performance made on the evaluation set are realistic, or even pessimistic, since the evaluation set is characterized by a new topic: Covid-19. For the evaluation set about 10,000 tweets were selected.

3.3.3 Annotation results

Each tweet was annotated twice: In 90% of the cases by two different annotators and in 10% of the cases by the same annotator. Special attention was devoted to evening out the overlap between annotators to get agreement estimates on equally sized sets.

Ten annotators were engaged for our annotation campaign. They were given annotation guidelines, a training session and a test on a small set to evaluate their understanding of the task and their commitment before starting the annotation procedure. The annotation process lasted four months, and it required about 1,200 person-hours for the ten annotators to complete the task.

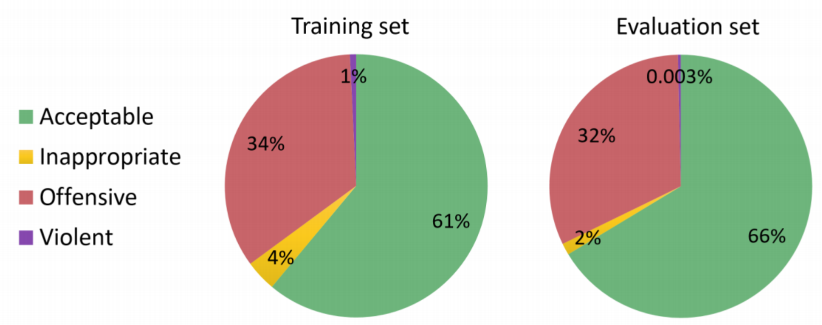


Figure 6: Distribution of types of hate speech in Slovenian Twitter datasets: on the training (left) and evaluation sets (right). The distributions differ, as the sampling for the training set was intentionally biased to contain more unacceptable speech. The evaluation set represents a random sample, therefore its proportion of violent hate speech is drastically smaller.

The annotation results for the Slovenian training and evaluation sets are summarized in Table 7.

Training set		**Evaluation set**	
Annotated for Vrsta:	99809	**Annotated for Vrsta:**	20000
0 ni sporni govor	60981	0 ni sporni govor	13273
1 nespodobni govor	3817	1 nespodobni govor	285
2 žalitev	34244	2 žalitev	6373
3 nasilje	767	3 nasilje	69
Annotated for Tarča:	34204	**Annotated for Tarča:**	6430
1 ksenofobija in rasizem	1103	1 ksenofobija in rasizem	125
2 begunci/migranti	1011	2 begunci/migranti	68
3 islamofobija	527	3 islamofobija	21
4 antisemitizem	55	4 antisemitizem	10
5 druge religije	172	5 druge religije	15
6 homofobija	304	6 homofobija	16
7 seksizem	773	7 seksizem	68
8 ideologija	6231	8 ideologija	839
9 novinarji in mediji	2517	9 novinarji in mediji	682
10 politika/-i	10924	10 politika/-i	2623
11 posameznik	7016	11 posameznik	1318
12 drugo	3571	12 drugo	645

Table 7: The numbers of different types (slo. vrsta) and targets (slo. tarča) of hate speech in the Slovenian training and evaluation datasets.

3.4 Dutch Data

We collected Dutch Twitter data from January 2018 until October 1 2020 using the official Twitter API to collect tweets for a wide variety of keywords, grouped in thematic and regional clusters. This resulted in a data set of about 16 million tweets. We also mined 3.4 million comments from Facebook groups active in the aforementioned thematic and regional clusters.

The collected tweets and comments were processed using the Textgain text analytics API to extract metadata features such as named entities and demographic features. A preliminary toxicity score and toxicity dimensions were applied using the Dutch POW-lexicon method[3]. For annotation, we preselected the 7000 most toxic records in order to ascertain in-domain data.

We additionally collected 19,000 comments from 300,000 YouTube videos in the aforementioned clusters and randomly selected 8,500 comments for annotation. We also collected 17,500 argument pairs from the internet forums geenstijl.nl and dumpert.nl, known for its recalcitrant and often misogynistic rhetoric and selected 13,000 for annotation.

3.4.1 Annotation

Each record was annotated by at least 2 annotators from a pool of 15 annotators. Annotation was done through an in-house TextGain annotation tool called Oncilla that monitors annotation speed, personal label distribution (Figure 7 and 8) and establishes systematic coupling of annotators. In case of disagreement, a 3rd or 4th annotator was asked to label the record with the aim of establishing a tie break.

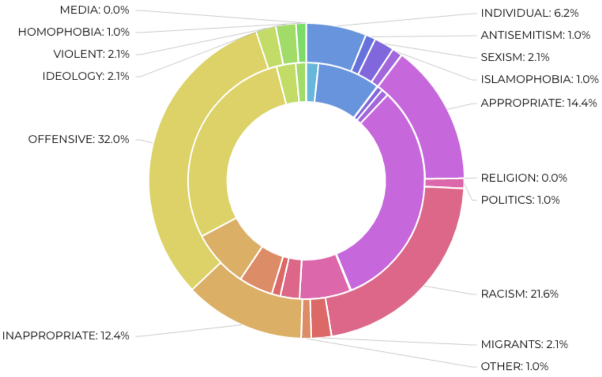
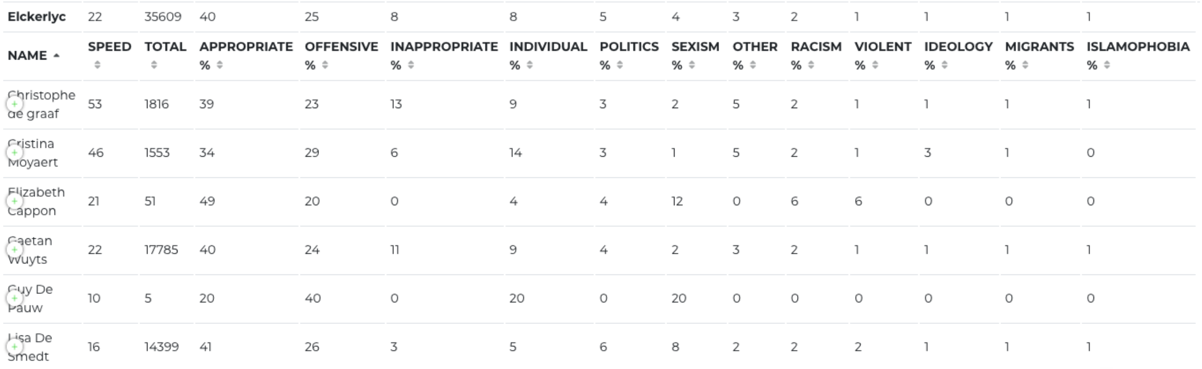


Figure 7:Self-assessment for the annotator to compare one's own label distribution (outer circle) vs the average label distribution of all annotators (inner circle).



Elckerlyc	22	35609	40	25	8	8	5	4	3	2	1	1	1	1
NAME ▲	**SPEED** ⇕	**TOTAL** ⇕	**APPROPRIATE %** ⇕	**OFFENSIVE %** ⇕	**INAPPROPRIATE %** ⇕	**INDIVIDUAL %** ⇕	**POLITICS %** ⇕	**SEXISM %** ⇕	**OTHER %** ⇕	**RACISM %** ⇕	**VIOLENT %** ⇕	**IDEOLOGY %** ⇕	**MIGRANTS %** ⇕	**ISLAMOPHOBIA %** ⇕
Christophe de graaf	53	1816	39	23	13	9	3	2	5	2	1	1	1	1
Cristina Moyaert	46	1553	34	29	6	14	3	1	5	2	1	3	1	0
Elizabeth Cappon	21	51	49	20	0	4	4	12	0	6	6	0	0	0
Gaetan Wuyts	22	17785	40	24	11	9	4	2	3	2	1	1	1	1
Guy De Pauw	10	5	20	40	0	20	0	20	0	0	0	0	0	0
Lisa De Smedt	16	14399	41	26	3	5	6	8	2	2	2	1	1	1

Figure 8 Admin monitoring function to inspect the team's annotation behavior.

[3] https://www.textgain.com/portfolio/profanity-offensive-words/

Around 29,000 records are annotated to date. Figure 99 shows the distribution of the labels for the entire dataset. Note the larger proportion of non-appropriate data compared to the other languages. This is due to the selection of records, which was based on (1) scores obtained from a profanity lexicon for tweets and Facebook comments, (2) data from thematic clusters, many of which tend to elicit hateful comments and (3) data from forums known for above-average toxic rhetoric. We will interface with the project partners to establish an evaluation set that is reflective of the expected distribution of real-life data.

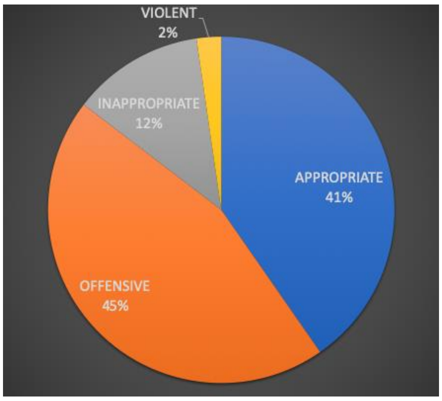


Figure 9: Distribution of types of hate speech in the Dutch data

3.4.2 Stance detection & level of disagreement

We also performed two additional annotation tasks that are relevant to the automated detection of hate speech: 14,000 documents were additionally annotated for stance detection, using the RumourEval 2017 annotation scheme (containing the labels: DENY – SUPPORT – QUERY – COMMENT). We also annotated 17,000 documents for level of disagreement using the taxonomy put forward by Paul Graham[4].

4 Acknowledgement

We would like to thank all the annotators who have collaborated in our data annotation campaigns. The following annotators agreed to disclose their names: Lucija Mandić, Katarina Bogataj, Jošt Žagar, Anna Maria Grego, Haris Agovič, Predrag Petrovič, Batisti Filippo, Galeazzi Alessandro, Gamannossi degl'Innocenti Matilde, Minelli Eleonora, Paone Antonietta, Sciumbata Floriana, Vidali Andrew, Vivan Elena.

[4] http://www.paulgraham.com/disagree.html

5 Appendix

5.1 Annotation guidelines

WP2 – Hate speech detection and trends

HATE SPEECH ANNOTATION GUIDELINES - English

June, 2020

AUTHORS: Igor Mozetič, Ajda Šulc, Petra Kralj Novak
Jožef Stefan Institute

This work was funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, grant no. 875263).

The content of this work represents the views of the authors only and is their sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

1. Introduction

IMSyPP – Innovative Monitoring Systems and Prevention Policies of Online Hate Speech – is a European Union's Rights, Equality and Citizenship Programme (2014-2020) action grant project (project ID 875263). IMSyPP is tackling hate speech in a multidisciplinary fashion combining machine learning, computational social science and linguistic approaches to support a data-driven approach to hate speech regulation, prevention and awareness-raising.

The goal of IMSyPP annotation campaigns is to label data that will be used for training machine learning classifiers, as one of the IMSyPP goals is automated detection and sustainable monitoring of hate speech. Therefore, we need to develop near real-time hate speech detection models tuned to language, culture and legislation, taking into account the context of the message. The data collected in this annotation campaign will be mainly used for training the hate speech detection models. In addition, it will allow us to assess how difficult/subjective detection of hate hate speech is.

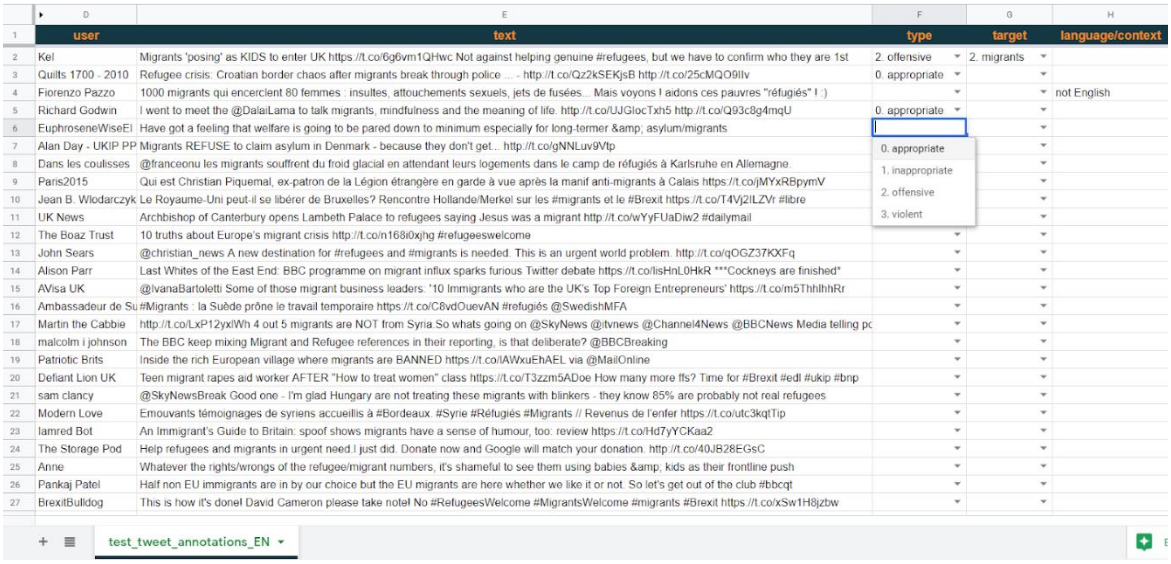
In IMSyPP, we are tackling user generated on-line text in several languages (English, Italian, Dutch and Slovenian) and several types of user generated content (tweets, YouTube & Facebook comments, comments on news sites, 4Chan posts). The posts should be annotated for type (appropriate, inappropriate, offensive, violent) and target (racism, migrants, islamophobia, antisemitism, religion, homophobia, sexism, ideology, media, politics, individual, other).

2. Annotation Interface

GoogleSheets is used as a user interface for annotations. The text of the comment (or tweet) is displayed in individual lines in the Google spreadsheet. On the right hand-side of the text, the annotator selects the appropriate categories on two levels: the type of discourse (appropriate to violent) and the target of any hate speech.

At this stage of the project, the context is deliberately not taken into account. Tweets are exported individually – the annotators (you) should treat them as unrelated tweets, without looking at the previous posts to which they respond or relate. In the case it is really not possible to determine the type or target of the post without the context, you should enter "context" in the last column. You should not search for tweets on the portal or through Google search to make sure of the context.

A screenshot of the annotation interface for tweets is depicted in Figure 1.



	D	E	F	G	H
1	user	text	type	target	language/context
2	Kel	Migrants 'posing' as KIDS to enter UK https://t.co/6g6vm1QHwc Not against helping genuine #refugees, but we have to confirm who they are 1st	2. offensive	2. migrants	
3	Quilts 1700 - 2010	Refugee crisis: Croatian border chaos after migrants break through police ... - http://t.co/Qz2kSEKjsB http://t.co/25cMQO9Ilv	0. appropriate		
4	Fiorenzo Pazzo	1000 migrants qui encerclent 80 femmes : insultes, attouchements sexuels, jets de fusées... Mais voyons ! aidons ces pauvres "réfugiés" ! :)			not English
5	Richard Godwin	I went to meet the @DalaiLama to talk migrants, mindfulness and the meaning of life. http://t.co/UJGlocTxh5 http://t.co/Q93c8g4mqU	0. appropriate		
6	EuphroseneWiseEl	Have got a feeling that welfare is going to be pared down to minimum especially for long-termer &amp; asylum/migrants			
7	Alan Day - UKIP PP	Migrants REFUSE to claim asylum in Denmark - because they don't get... http://t.co/gNNLuv9Vtp			
8	Dans les coulisses	@franceonu les migrants souffrent du froid glacial en attendant leurs logements dans le camp de réfugiés à Karlsruhe en Allemagne.			
9	Paris2015	Qui est Christian Piquemal, ex-patron de la Légion étrangère en garde à vue après la manif anti-migrants à Calais https://t.co/jMYxRBpymV			
10	Jean B. Wlodarczyk	Le Royaume-Uni peut-il se libérer de Bruxelles? Rencontre Hollande/Merkel sur les #migrants et le #Brexit https://t.co/T4Vj2ILZVr #libre			
11	UK News	Archbishop of Canterbury opens Lambeth Palace to refugees saying Jesus was a migrant http://t.co/wYyFUaDiw2 #dailymail			
12	The Boaz Trust	10 truths about Europe's migrant crisis http://t.co/n168i0xjhg #refugeeswelcome			
13	John Sears	@christian_news A new destination for #refugees and #migrants is needed. This is an urgent world problem. http://t.co/qOGZ37KXFq			
14	Alison Parr	Last Whites of the East End: BBC programme on migrant influx sparks furious Twitter debate https://t.co/lisHnL0HkR ***Cockneys are finished*			
15	AVisa UK	@IvanaBartoletti Some of those migrant business leaders: '10 Immigrants who are the UK's Top Foreign Entrepreneurs' https://t.co/m5ThhlhhRr			
16	Ambassadeur de Su	#Migrants : la Suède prône le travail temporaire https://t.co/C8vdOuevAN #refugiés @SwedishMFA			
17	Martin the Cabbie	http://t.co/LxP12yxlWh 4 out 5 migrants are NOT from Syria.So whats going on @SkyNews @itvnews @Channel4News @BBCNews Media telling pc			
18	malcolm i johnson	The BBC keep mixing Migrant and Refugee references in their reporting, is that deliberate? @BBCBreaking			
19	Patriotic Brits	Inside the rich European village where migrants are BANNED https://t.co/lAWxuEhAEL via @MailOnline			
20	Defiant Lion UK	Teen migrant rapes aid worker AFTER "How to treat women" class https://t.co/T3zzm5ADoe How many more ffs? Time for #Brexit #edl #ukip #bnp			
21	sam clancy	@SkyNewsBreak Good one - I'm glad Hungary are not treating these migrants with blinkers - they know 85% are probably not real refugees			
22	Modern Love	Emouvants témoignages de syriens accueillis à #Bordeaux. #Syrie #Réfugiés #Migrants // Revenus de l'enfer https://t.co/utc3kqtTip			
23	Iamred Bot	An Immigrant's Guide to Britain: spoof shows migrants have a sense of humour, too: review https://t.co/Hd7yYCKaa2			
24	The Storage Pod	Help refugees and migrants in urgent need.I just did. Donate now and Google will match your donation. http://t.co/40JB28EGsC			
25	Anne	Whatever the rights/wrongs of the refugee/migrant numbers, it's shameful to see them using babies &amp; kids as their frontline push			
26	Pankaj Patel	Half non EU immigrants are in by our choice but the EU migrants are here whether we like it or not. So let's get out of the club #bbcqt			
27	BrexitBulldog	This is how it's done! David Cameron please take note! No #RefugeesWelcome #MigrantsWelcome #migrants #Brexit https://t.co/xSw1H8jzbw			

Figure 1: Screenshot of the IMSyPP annotation interface showing a drop-down menu.

YouTube and Facebook comments are listed in threads. A “-*-*-*-” at the beginning of the line denotes a reply to a previous comment (above without “-*-*-*-”). There can be several replies to a single comment. Comments are chronologically ordered.

In the annotation interface, drop-down menus of the possible categories (labels) are encoded with numbers (same as in the two lists below). These allow:

- By pressing "Enter" or mouse clicking on the selected cell, a drop-down menu with all possible labels will be displayed.
- Selecting the appropriate label by entering part of the word (type "Insu…", "Vio…" in the cell, the table automatically suggests a category that corresponds to what is displayed (Insult, Violence), press "Enter" to confirm),
- Selecting the appropriate label by the corresponding number (type "0" in the cell and press "Enter", the table automatically displays the category "0 Appropriate speech".

3. Hate speech type

At the speech type level, you can choose between four categories:

1. Appropriate - no target (leave the "target" category blank)
2. Inappropriate (contains terms that are obscene, vulgar; but the text is not directed at any person specifically) - has no target (leave the “target” category blank)
3. Offensive (including offensive generalization, contempt, dehumanization, indirect offensive remarks)
4. Violent (author threatens, indulges, desires or calls for physical violence against a target; it also includes calling for, denying or glorifying war crimes and crimes against humanity)

If the post contains several different types of unacceptable discourse, select the type the highest in the hierarchy (1 < 2 < 3).

In the case of quoted hate speech, consider the intention of the author. If it is a reproduction and agreement with offensive content, mark it as "insulting". If it is a quote and a critique of hostility, mark it with “appropriate” (in case the critique does not contain offensive or obscene terms).

4. Hate speech target

At the level of the target of hate speech, you can choose between 12 categories:

1. Racism (intolerance based on nationality, ethnicity, language, towards foreigners; and based on race, skin color)
2. Migrants (intolerance of refugees or migrants, offensive generalization, call for their exclusion, restriction of rights, non-acceptance, denial of assistance…)
3. Islamophobia (intolerance towards Muslims)
4. Antisemitism (intolerance of Jews; also includes conspiracy theories, Holocaust denial or glorification, offensive stereotypes…)
5. Religion (other than above)
6. Homophobia (intolerance based on sexual orientation and / or identity, calls for restrictions on the rights of LGBTQ persons
7. Sexism (offensive gender-based generalization, misogynistic insults, unjustified gender discrimination)
8. Ideology (intolerance based on political affiliation, political belief, ideology… e.g. “communists”, “leftists”, “home defenders”, “socialists”, “activists for…”)
9. Media (journalists and media, also includes allegations of unprofessional reporting, false news, bias)
10. Politics (intolerance towards individual politicians, authorities, system, political parties)

11. Individual (intolerance toward any other individual due to individual characteristics; like commentator, neighbor, acquaintance)
12. Other (intolerance towards members of other groups due to belonging to this group; write in the blank column on the right which group it is)

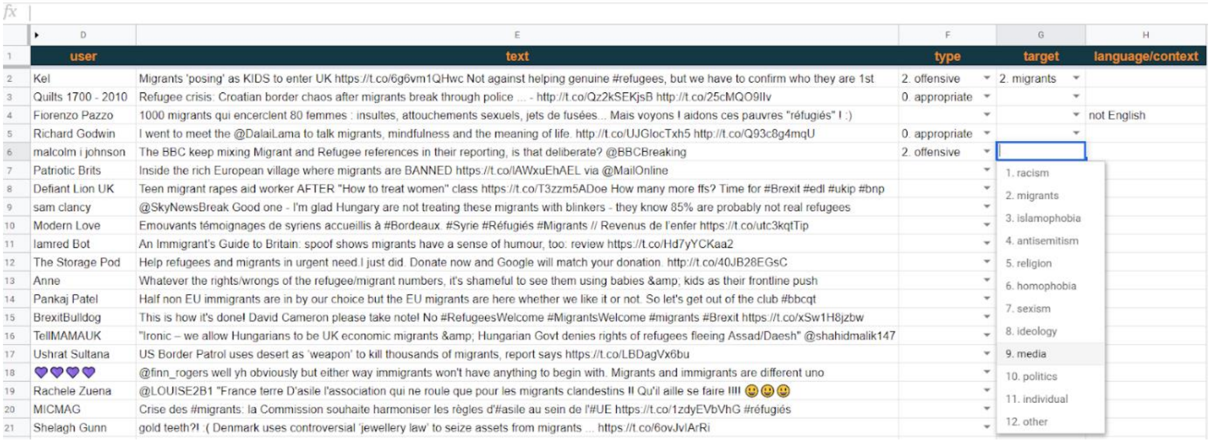
If the target itself can be classified into several categories, indicate the one for which it is targeted. Examples:

- If the post is about refugees but primarily insults them for belonging to Islam, it is Islamophobia;
- If it offends the Catholic Church members as individuals, not for their catholic affiliation, e.g. certain pastors as pedophiles, but does not generalize this to all catholic believers, it is "other" (if it generalizes to pastors) or "individual" (if it offends only some individual priests), it is not "other religions" in the sense of insulting Christian believers).

If a post contains several different targets of hate speech:

a. Select a target against a hierarchically higher type of hate speech (violence), or
b. If the type is the same for all targets, select the target to which the text is most offensive.

A screenshot of the annotation interface depicting the drop-down for selecting hate speech target is depicted in Figure 2.



1	user	text	type	target	language/context
2	Kel	Migrants 'posing' as KIDS to enter UK https://t.co/6g6vm1QHwc Not against helping genuine #refugees, but we have to confirm who they are 1st	2. offensive	2. migrants	
3	Quilts 1700 - 2010	Refugee crisis: Croatian border chaos after migrants break through police ... - http://t.co/Qz2kSEKjsB http://t.co/25cMQO9Ilv	0. appropriate		
4	Fiorenzo Pazzo	1000 migrants qui encerclent 80 femmes : insultes, attouchements sexuels, jets de fusées... Mais voyons ! aidons ces pauvres "réfugiés" ! :)			not English
5	Richard Godwin	I went to meet the @DalaiLama to talk migrants, mindfulness and the meaning of life. http://t.co/UJGlocTxh5 http://t.co/Q93c8g4mqU	0. appropriate		
6	malcolm i johnson	The BBC keep mixing Migrant and Refugee references in their reporting, is that deliberate? @BBCBreaking	2. offensive		
7	Patriotic Brits	Inside the rich European village where migrants are BANNED https://t.co/lAWxuEhAEL via @MailOnline			
8	Defiant Lion UK	Teen migrant rapes aid worker AFTER "How to treat women" class https://t.co/T3zzm5ADoe How many more ffs? Time for #Brexit #edl #ukip #bnp			
9	sam clancy	@SkyNewsBreak Good one - I'm glad Hungary are not treating these migrants with blinkers - they know 85% are probably not real refugees			
10	Modern Love	Emouvants témoignages de syriens accueillis à #Bordeaux. #Syrie #Réfugiés #Migrants // Revenus de l'enfer https://t.co/utc3kqtTip			
11	Iamred Bot	An Immigrant's Guide to Britain: spoof shows migrants have a sense of humour, too: review https://t.co/Hd7yYCKaa2			
12	The Storage Pod	Help refugees and migrants in urgent need.I just did. Donate now and Google will match your donation. http://t.co/40JB28EGsC			
13	Anne	Whatever the rights/wrongs of the refugee/migrant numbers, it's shameful to see them using babies &amp; kids as their frontline push			
14	Pankaj Patel	Half non EU immigrants are in by our choice but the EU migrants are here whether we like it or not. So let's get out of the club #bbcqt			
15	BrexitBulldog	This is how it's done! David Cameron please take note! No #RefugeesWelcome #MigrantsWelcome #migrants #Brexit https://t.co/xSw1H8jzbw			
16	TellMAMAUK	"Ironic – we allow Hungarians to be UK economic migrants &amp; Hungarian Govt denies rights of refugees fleeing Assad/Daesh" @shahidmalik147			
17	Ushrat Sultana	US Border Patrol uses desert as 'weapon' to kill thousands of migrants, report says https://t.co/LBDagVx6bu			
18	💜💜💜💜	@finn_rogers well yh obviously but either way immigrants won't have anything to begin with. Migrants and immigrants are different uno			
19	Rachele Zuena	@LOUISE2B1 "France terre D'asile l'association qui ne roule que pour les migrants clandestins !! Qu'il aille se faire !!!! 😀😀😀			
20	MICMAG	Crise des #migrants: la Commission souhaite harmoniser les règles d'#asile au sein de l'#UE https://t.co/1zdyEVbVhG #réfugiés			
21	Shelagh Gunn	gold teeth?! :(Denmark uses controversial 'jewellery law' to seize assets from migrants ... https://t.co/6ovJvIArRi			

Figure 2 Screenshot depicting the drop-down menu for selecting the Target of hate speech.

5. Other guidelines

Do not open links in the posts that contain links to other sites.

If the text is written in another language, enter "other language" in the last column.

Even if you, as an annotator, agree with the written insult or negative criticism (e.g., in the media - about bias, about fake news), such a post should still be marked with an appropriate label of hate speech (as "insult" or "violence").