EVALITA

Evaluation of NLP and Speech Tools for Italian

HaSpeeDe 3

Political and Religious Hate Speech Detection

Mirko Lai, Fabio Celli, Alan Ramponi,
Sara Tonelli, Cristina Bosco, Viviana Patti

Parma, September 7th-8th 2023

AILC Associazione Italiana di Linguistica Computazionale

Motivations

Online hateful content, or Hate Speech (HS) could be equally or more dangerous than offline communications.

Hate Speech is a proxy for global increase in violence toward minorities and detect it is crucial to preserve free speech and democracy.

Therefore, its automatic identification has become a crucial mission in many fields.

M. L. Williams at al., ***Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime****,*
The British Journal of Criminology, Volume 60, Issue 1, January 2020, Pages 93–117
https://doi.org/10.1093/bjc/azz049

Motivations

Computational Ethics

HaSpeeDe 3 – Political and Religious Hate Speech Detection

HODI – Homotransphobia Detection in Italian

MULTI-Fake-DetectiVE – MULTImodal Fake News Detection and VErification

ACTI – Automatic Conspiracy Theory Identification

B. Chulvi et al. ***Fake News and Hate Speech: Language in Common****,*
arXiv e-prints, 2022. *https://doi.org/10.48550/arXiv.2212.02352*

Previous shared tasks

HaSpeeDe

HaSpeeDe 2

HaSpeeDe and HaSpeeDe 2 focused on HS against **immigrants**, **Muslims** and **Roms**;

HaSpeeDe 3 explores HS in strong polarised debates, in particular concerning **political** and **religious** topics.

HaSpeeDe 3

Previous shared tasks

EVALITA 2020

Sardistance (stance detection)

Similarly to **Sardistance**, **HaSpeeDe 3** paid attention on **contextual information** about the **authors** of the tweets

Data Collection

data collection from Twitter

(that changed its logo in the meanwhile) 🤔

Using both the **Stream API** and the **API v2 for academic research**.

👋 **APIs not freely available anymore** 👋

Datasets

PolicyCorpusXL

The dataset contains **7000 tweets** collected employing a snowball sampling from three starting hashtags (#dpcm, #legge, #leggedibilancio). 5736 tweets have been collected between April and July 2021 and 1264 between March and May 2020

ReligiousHate

The dataset is composed of **3000 tweets** collected between December 2020 and August 2021 with keywords that refer to the three main monotheistic religions, namely Christianity, Islam and Judaism

Annotation Schema

a binary classification problem

HS	¬HS
the tweet contains hatred	the tweet doesn't contain hatred
❗	✓

Annotation

PolicyCorpusXL
(Fleiss' k = 0.53)

- 2 annotators annotated the entire dataset.
- a third annotation in case of disagreement.
- 1000 tweets have been finally discarded for artificially augmenting the portion of hate tweets.

ReligiousHate
(Cohen's k = 0.57)

- 3 native speakers of Italian w/ background in linguistics and computer science;
- Protocol that foresaw in-person discussion rounds and adjudication sessions

Label distribution

Set	Dataset	HS	¬HS	Total
dev set	*PolicyCorpusXL*	3456 ~62%	2144 ~38%	5600 100%
	ReligiousHate	-	-	-
test set	*PolicyCorpusXL*	700 50%	700 50%	1400 100%
	ReligiousHate	487 ~16%	2513 ~84%	3000 100%
	Total	4643	5357	10000

Data Format

textual information

- **anonymized_tweet_id**: A pseudo-random integer that identifies the specific tweet and replaces the original tweet id
- **anonymized_text**
  - URLs have been replaced by the placeholder [URL]
  - mentions have been replaced and mapped by a pseudo-random integer that identifies a specific user.
- **label**: 1 for hateful tweets, 0 otherwise.
- **dataset**: this field specifies the set (training or test) and whether a tweet belongs to the PolicyCorpusXL or the ReligiousHate dataset.

Data Format

contextual information (about the **tweet**)

- **created_at**: The posting date of the tweet.
- **retweet_count**: The number of times the tweet has been retweeted.
- **favorite_count**: It indicates approximately how many times this tweet has been liked by Twitter users.
- **source**: The source used for posting the tweet (e.g., Android, iOS).
- **is_reply**: 1 if the tweet is a reply, 0 otherwise.
- **is_retweet**: 1 if the tweet is a retweet, 0 otherwise.
- **is_quote**: 1 if the tweet is a quote, 0 otherwise.

Data Format

contextual information (about the **user**)

- **anonymized_user_id**: The original author id (if known), replaced by a pseudo-random integer.
- **user_created_at**: The date when the author created the account.
- **statuses_count**: The number of tweets posted by the author.
- **followers_count:** The number of Twitter users that follow the author.
- **friends_count**: The number of Twitter users that the author follows.
- **anonymized_description**: The self-description of the author of the tweet. We applied the same anonymisation strategy applied to the field anonymized_text.

Data Format

contextual information (about the **user social network**)

- **friendship relations**:
  - **source**: A user, identified by anonymized_user_id, that follows the target
  - **target**: A user, identified by anonymized_user_id, that is followed by the source.
- **retweet/reply relations**:
  - **source**: A user, identified by anonymized_user_id, that retweeted target
  - **target**: A user, identified by anonymized_user_id, that has been retweeted/replied by source.
  - **date**: The day when source retweeted/replied target.
  - **count**: The number of times the source retweeted/replied the target that day.

Definition of the Task

Task A – (constrained in-domain) political hate speech detection:
a binary classification task aimed at determining whether a message contains hate speech or not. The task is based on the PolicyCorpusXL dataset.

It comprises the following subtasks:

– **Textual only**: participants can only use the provided textual content of the tweets from PolicyCorpusXL for development;
– **Textual+Contextual**: participants can employ for development the textual content of the tweets plus contextual information given to them (i.e., metadata of the tweet and author, friends, retweets, and reply relations).

Definition of the Task

Task B – (unconstrained) Cross-domain hate speech detection:
a binary classification task with test data from different domains – i.e., political and religious. The main objective of this task is to explore cross-domain hate speech detection under two evaluation settings:

– **XPoliticalHate**: the test set consists of tweets from PolicyCorpusXL;
– **XReligiousHate**: the test set consists of tweets from the ReligiousHate corpus, for which no development data is provided to participants.

Moreover, participants are allowed to use **any kind of external data** (e.g., datasets for other hate domains) and textual and contextual PolicyCorpusXL development data.

Evaluation Metrics

We provide **four separate official rankings**, one for each subtask.
Participants can submit **two runs** for each subtask.
However, participants are not required to participate in all subtasks or to submit 2 runs for each of them.

Submissions are ranked by averaged F_1- score over the two classes, according to the following equation:

$$F_1(avg) = (F_1^{HS} + F_1^{\neg HS})/2$$

6 Participants

BERTicelli: Antwerp, **Belgium**

CHILab: Palermo, **Italy**

extremITA: Rome “Tor Vergata”, Turin, **Italy**

INGEOTEC: Aguascalientes, Ciudad de México, **México**

LMU: Munich, **Germany**

odang4: London, **United Kingdom**, Bologna, Turin, **Italy**

Overall results

- No teams benefited from contextual information, few teams employed them

- No teams benefited from external data sources dealing with **Task B: XPoliticalHate**

- Few teams benefited from external data sources dealing with **Task B: XReligiousHate**

- All teams benefited from pre-trained language model (e.g. ALBERTo, UmBERTo, LLaMA)

Final ranking

Team	Task A (in-domain | political)				Task B (cross-domain)			
	textual		contextual		XPoliticalHate		XReligiousHate	
	run 1	run 2	run 1	run 2	run 1	run 2	run 1	run 2
odang4	**0.9128**	0.8950	**0.9128**	0.8950	**0.9128**	0.8950	0.5213	0.4809
extremITA	0.9079	0.9034	0.9079	0.9034	0.9079	0.9034	0.5921	**0.6525**
LMU					0.9014	0.8984	0.6458	0.6461
BERTicelli	0.8976	0.8652	0.8976	0.8969	0.8976	0.8969	0.5401	0.5384
INGEOTEC	0.8845		0.8845		0.8845		0.5522	
CHILab	0.8257	0.8516	0.8257	0.8516	0.8257	0.8516		
avg		0.8826		0.8862		0.8887		0.5744
std		0.0293		0.0288		0.0264		0.0624

Thank you

Any Questions?