HaSpeeDe3 at EVALITA 2023: Overview of the Political and Religious Hate Speech Detection task

Mirko Lai[1,*], Fabio Celli[2], Alan Ramponi[3], Sara Tonelli[3], Cristina Bosco[1] and Viviana Patti[1]

[1]*Università degli Studi di Torino, Torino, Italy*
[2]*Maggioli s.p.a., University of Trento, Trento, Italy*
[3]*Fondazione Bruno Kessler (FBK), Trento, Italy*

Abstract

The Hate Speech Detection (HaSpeeDe3) task is the third edition of a shared task on the detection of hateful content in Italian tweets. It differs from the previous editions while maintaining continuity in analysing and contrasting hate speech (HS) on social media. While HaSpeeDe and HaSpeeDe2 were focused on HS against immigrants, Muslims and Roms, HaSpeeDe3 explores hate speech in strong polarised debates, concerning in particular politics and religion. It is articulated in two different tasks: A) In-domain political hate speech detection and B) Cross-domain hate speech detection about political and religious tweets. Task A consists in two different subtasks for which participants i) can only use the provided textual content of the tweet, or ii) can additionally employ contextual information about the tweet and its author. In Task B, that consists in two subtasks, participants are allowed to use any kind of external data for detecting hate speech in tweets about i) politics and ii) religion. Six teams from both academia and industry participated in the evaluation, with a total of 13 submitted runs for Task A and 16 for Task B.

Keywords

Hate speech detection, social media analysis, polarised debates, political hate speech, religious hate speech, shared task

1. Introduction and Motivation

Social media play an important role in public debates, especially concerning politics. On the one hand, political leaders use social media as a vehicle for political and electoral propaganda. On the other hand, they provide news to a significant part of the population that takes part in the discussion, supporting or criticising political decisions [1, 2]. Social media are also the place where debates on sensitive topics, such as religious beliefs and practices, are rather common and sometimes are intertwined with public discussions on political matters.

Unfortunately, such discussions often trigger verbal aggressions [3], especially after some polarising events in Europe and beyond such as Brexit [4], the Covid-19 pandemics [5] and the Russo-Ukrainian conflict [6]. Aggressions and online hate are exacerbated by the ideological segregation present on social media, where social homophily, as well as personalising and recommending algorithms, facilitate the creation of *echo chambers* and *filter bubbles* [7, 8]. The "others" are frequently targeted because of characteristics such as gender, sexual orientation, ethnicity, and religion [9, 10, 11].

In the last years, to address these problems posed by the widespread use of abusive language online, the NLP community has focused on the detection of hate speech [12] and the analysis of online debates [13, 14]. In particular, many researchers have worked on systems to detect offensive language against specific vulnerable groups, e.g., women, immigrants, LGBTQ+ community, among others [11, 15, 16, 17]. An under-researched – yet important – area of investigation is anti-politics hate, i.e., hate speech against politicians, policy makers and laws at any level (national, regional and local). While anti-policy hate speech has been addressed in Arabic [18] and German [19], most European languages have been under-researched. As regards religious hate, instead, annotated corpora have been created for English, Arabic, Bengali, French, Portuguese, and Italian, among others (for an overview of works, see [15] and [20]). However, none of them share contextual information about the authors of the tweets, neither about their social media network, although religious self-identification may lead to hard conflict with the members of other worships.

For this shared task organised within EVALITA 2023 [21], we introduce a new corpus, called PolicyCorpusXL, containing Italian tweets related to political topics, where

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
mirko.lai@unito.it (M. Lai); Fabio.Celli@maggioli.it (F. Celli); alramponi@fbk.eu (A. Ramponi); satonelli@fbk.eu (S. Tonelli); cristina.bosco@unito.it (C. Bosco); viviana.patti@unito.it (V. Patti)
http://www.di.unito.it/~lai/ (M. Lai); https://dh.fbk.eu/author/alramponi/ (A. Ramponi); https://dh.fbk.eu/author/sara/ (S. Tonelli); http://www.di.unito.it/~bosco/ (C. Bosco); http://www.di.unito.it/~patti/ (V. Patti)
0000-0003-1042-0861 (M. Lai); 0000-0002-7309-5886 (F. Celli); 0000-0002-4305-2404 (A. Ramponi); 0000-0001-8010-6689 (S. Tonelli); 0000-0002-8857-4484 (C. Bosco); 0000-0001-5991-370X (V. Patti)

hateful messages have been manually annotated. This corpus is an extension of PolicyCorpus [22]. We selected Twitter as the source of data and Italian as the target language because Italy has, at least since the elections in 2018, a large audience that pays attention to hyper-partisan sources on Twitter. These users are prone to produce and retweet messages of hate against policy-making [23]. We also provide the Italian portion of the ReligiousHate dataset [20] as a test set, in which hateful tweets concerning Christianity, Islam and Judaism have been manually labeled. Our goal is to test the in-domain performance of systems for political hate speech detection, as well as the out-of-domain performance on a test set about religion.

2. Definition of the Task

HaSpeeDe3 focuses on detecting hate speech in strong polarised debates on social media, in particular debates on Twitter about political and religious topics. With this task, we invite participants to explore not only features based on the textual content of the tweet, but also features based on contextual information such as metadata that describe both the tweet and the author, or information about the social media community of the participants of the debate.

We propose two tasks, A and B, that in the rest of the paper will be referred also as in-domain and cross-domain tracks. Both tasks aim at tackling binary classification problems, and thus participants' systems have to predict whether a tweet contains hatred or not. Each task consists of two subtasks:

- **Task A – (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 (Section 3) and comprises the following subtasks:
  - **Textual**: participants can only use the provided textual content of the tweets from PolicyCorpusXL for development;
  - **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).
- **Task B – 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 (as in both the in-domain subtasks above);
  - **XReligiousHate**: the test set consists of tweets from the ReligiousHate corpus (Section 3), 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.

3. Dataset and Format

In this section, we describe the dataset creation process (Section 3.1), including data collection, annotation, enrichment, and label distribution. Then, we outline the format used for sharing data to participants (Section 3.2).

3.1. Dataset Creation

We collected data from Twitter after selecting it among existing social media platforms where hatred content could be present. There are two main reasons for this choice. One the one hand, Twitter easily allows the retrieval of a high volume of textual content by using APIs. On the other hand, additional metadata about tweets themselves and their authors can be collected. Furthermore, Twitter users can perform asynchronous actions such as retweeting, replying, and following. This latter aspect allows us to share with HaSpeeDe3 participants not only the text of the tweets and their metadata but also contextual information about the network where the participants of the online debate are situated.

3.1.1. Data Collection

The focus of HaSpeeDe3 is the detection of hate speech in strong polarised debates, in particular concerning political and religious topics. We use two different datasets for the shared task: **PolicyCorpusXL** [24] and **ReligiousHate** [20]. For both of them, tweets have been collected via the Twitter APIs by querying for key terms specific to each topic.

PolicyCorpusXL The dataset contains 7,000 tweets collected employing a snowball sampling from three starting hashtags (#dpcm, #legge, #leggedibilancio). 5,736 tweets have been collected between April and July 2021 and 1,264 between March and May 2020 [22].

ReligiousHate We use the Italian portion of the religious hate speech corpus introduced in [20]. The dataset is composed of 3,000 tweets collected between December

2020 and August 2021 with keywords that refer to the three main monotheistic religions, namely Christianity, Islam and Judaism.

Due to the different nature of the political and religious topics, the protocols used for data collection are not the same; however, in both cases, offensive words have not been used as query terms to minimise biased dataset composition and potential learning shortcuts [25, 26].

3.1.2. Data Annotation

We summarise the annotation procedure followed for PolicyCorpusXL and ReligiousHate below.

PolicyCorpusXL Two Italian experts of communication annotated the entire dataset. The training set has been additionally annotated by a third expert in case of disagreement. 1,000 tweets have been finally discarded in order to artificially augment the portion of hate tweets and provide more information for the classifiers. With this strategy, the number of tweets containing hate increased from 11.8% (a typical percentage obtained with random sampling) to 40.6%.

ReligiousHate Three native speakers of Italian with a background in linguistics and computer science annotated 3,000 tweets about religion that have been collected as described in Section 3. Annotation was performed following a protocol for experts that foresaw in-person discussion rounds and adjudication sessions.

The Inter Annotation Agreement is similar for both the PolicyCorpusXL (Fleiss' $k = 0.53$) and ReligiousHate (Cohen's $k = 0.57$) datasets.

3.1.3. Data Enrichment

Using the Twitter stream APIs we retrieve tweets but we miss their subsequent history in the micro-blog platform. Indeed, since tweets are retrieved at posting time, we are not able to know what happens to them afterwards. In order to follow up the impact of a tweet on the user community after the posting time, we, therefore, use Twitter's APIs also to retrieve information about each tweet a posteriori. This makes it possible to check, for example, the number of times that the tweet has been retweeted or liked over the weeks after its posting time. We also collected a variety of additional information about the author, such as the list of friends and the users that each author retweeted and replied to since about 2018.

3.1.4. Label Distribution

Statistics of the two HaSpeeDe3's datasets are summarised in Table 1. PolicyCorpusXL consists of 7,000 tweets about political debates (5,600 in the development set and 1,400 in the test set), whereas ReligiousHate comprises 3,000 tweets, all belonging to the test set.

Table 1
Statistics of the datasets used for the HaSpeeDe3 shared task. HS: hate speech, ¬HS: non-hate speech.

Split	Dataset	HS	¬HS	Total
dev set	PolicyCorpusXL	3,456	2,144	5,600
	ReligiousHate	—	—	—
test set	PolicyCorpusXL	700	700	1,400
	ReligiousHate	487	2,513	3,000
	Total	4,643	5,357	10,000

3.2. Data Format

The development set consist of 5,600 tweets belonging to PolicyCorpusXL. It is organised in the following files, each including the following fields:

training|test_textual.csv

- 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]` and mentions have been replaced and mapped by a pseudo-random integer that identifies a specific user. For example, given the original content "*@MarioRossi Devi rinascere forse 100 volte per poter solo nominare il tuo incubo #Renzi*", the anonymised tweet is "*203951222958528 Devi rinascere forse 100 volte per poter solo nominare il tuo incubo #Renzi*". This means that the pseudo-random integer 203951222958528 identifies all forms of mentions of @MarioRossi, as the author of a tweet, and as a source or target in friends, retweets, and replies relations from/to @MarioRossi.
- label: `1` for hateful tweets, `0` otherwise. Test set labels have been released through the file test_textual_gold.csv after the competition ended.
- dataset: this field specifies the set (training or test) and whether a tweet belongs to the PolicyCorpusXL or the ReligiousHate dataset. training_textual.csv exclusively contains tweets belonging to the PolicyCorpusXL dataset.

training|test_contextual.csv

- anonymized_tweet_id: A pseudo-random integer that identifies the specific tweet and replaces the original tweet id.
- 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[1].
- source: The source used for posting the tweet (e.g., Android, iOS, web).
- 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.
- 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 of the file train_textual.csv described above.

The value of some fields could be unavailable or set to 0 if we were unable to recover the metadata of the tweet in 2022 (many months after the posting date), for example, because the tweet has been removed by Twitter, deleted, or made unavailable by the author.

training|test_contextual_friends.csv

- 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.

training|test_contextual_retweet|reply.csv

- source: A user, identified by anonymized_user_id, that retweeted target.
- target: A user, identified by anonymized_user_id, that has been retweeted by source.
- date: The day when source retweeted target.
- count: The number of times the source retweeted the target that day.

All sources are authors of at least one tweet in the training corpus, but some authors are missing in this file since it was not possible to recover their friend list.

All files described above are available at the official GitHub page of the task[2].

4. Evaluation Measures

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.

Systems are evaluated using F_1-score computed over the two binary classes, i.e., hate speech (HS) or non-hate speech (¬HS). Therefore, submissions are ranked by averaged F_1-score over the two classes, according to the following equation:

$$F_{1(avg)} = (F_1^{HS} + F_1^{\neg HS})/2$$

4.1. Baselines

We computed baselines using a simple machine learning model. For Task A - Textual, we employed a Support Vector Classifier trained with a unigram representation of the textual content of the tweet. For Task A - Contextual, we devise a baseline using the same classifier as above, based on a unigram representation of the textual content of the tweet, plus the number of retweets and favourites received by the tweet (retweet_count and favourite_count, see Section 3.2), the author degree computed from the friends network, and the author eigenvector centrality computed from the friends network. A last baseline for both the cross-domain hate speech subtasks employs a Support Vector Classifier with a unigram representation of the textual content of the tweet, trained with the XPoliticalHate and HaSpeeDe2 training sets [27].

In Table 2 we present the results obtained by the baselines on the four subtasks.

5. Task Overview: Participation and Results

A total of six teams participated in the HaSpeeDe3 task. We summarise their contribution below.

[1]Twitter released a number that "indicates approximately how many times th[e] Tweet has been liked by Twitter users": https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

[2]https://github.com/mirkolai/EVALITA2023-HaSpeeDe3

Table 2
Results obtained by the baselines on the four subtasks.

Task/subtask	$F_{1(avg)}$
Task A - Textual	0.8457
Task A - Contextual	0.8457
Task B - PoliticalHate	0.8458
Task B - ReligiousHate	0.5718

BERTicelli [28] The team submitted results for all the tasks and used all the provided sets of information. They exploited two pre-trained cased LLMs for Italian, namely UmBERTo and Italian BERT. In the pre-processing phase, they turned hashtags into words to reduce noise, they performed fine-tuning and used a 5-fold cross-validation for the Textual subtask, obtaining high scores. For the Contextual subtask, the team adopted an ensemble approach, wherein additional features were added to the fine-tuned models through a GradientBoosterClassifier algorithm. UmBERTo performed competitively in both Textual and Contextual subtasks but the model did not benefit from the addition of contextual features. Italian BERT, on the other hand, performed above the baselines but significantly lower than the task average. Overall, the team performed above the average in the political hate domain and below the average in the religious hate domain.

CHILab [29] The team participated only to the Task A - Textual, i.e., addressing only in-domain political hate speech detection using the provided textual content of the tweets from PolicyCorpusXL for development. They submitted two runs that employ two different models based on BiLSTM. The first one generates embeddings of 768 tokens from AlBERTo and the second one employs fastText for generating 300-dimensional token embeddings. Particular attention was paid to pre-processing. The `[URL]` tag, mention references, and retweet notes were removed since they were not considered relevant. Case sensitivity has been preserved as well as emojis due to the fact that they convey a specific meaning in social media communication in terms of prosody and emotions.

extremITA [30] The team addressed all the tasks using all the provided sets of information made available by the organisers. They also made use of data from all the EVALITA 2023 challenges to build monolithic architectures to tackle all the tasks. Their approaches are based on *i)* the IT5 encoder-decoder model, and *ii)* an instruction-tuned large language model built upon LLaMA. To the goal, for both models, they devised natural language instructions and output templates for each EVALITA task, including HaSpeeDe3. Among their submissions, we observe that the LLaMA-based model achieved better results than the IT5 one on Task A - Contextual, whereas the IT5 model achieved better results on the remainder subtasks.

INGEOTEC The team did not submit a system description report; therefore, we are unable to discuss and analyse their approach. They participated to the Task A - Textual and to the Task B - XReligiousHate considering both the evaluation settings.

LMU [31] The team participated only to the Task B - XReligiousHate considering both the evaluation settings with multitask prompt-training systems. Their systems consist of two steps in which models are *i)* pre-finetuned on external datasets in Italian and English from various domains, *ii)* fine-tuned on the target domain (only applicable to PolicyCorpusXL). As a backbone of their systems, they experimented with both Italian and multilingual pre-trained language models (PLMs). They showed that Italian datasets are more beneficial than the combination of Italian and English ones and that systems based on both Italian and multilingual PLMs achieved similar performance. Their best runs for the political and religious domains are ensembles of prompt-training systems based on Italian and multilingual PLMs.

odang4 [32] The team participated in both Tasks A and B, using only textual information in the former. They based their approach on the assumption that a relation between named entities and abusive language exists. They submitted two different runs. The first one employs enhanced-ALBERTo with triple verbalisation from the Ontology of Dangerous Speech [33] with prompting Davinci model. The second one applies a majority voting criteria among ALBERTo, the enhanced-ALBERTo with triple verbalisation from the Ontology of Dangerous Speech, and the enhanced-ALBERTo with prompting Davinci. For what concerns Task B - XReligiousHate, the multilingual expert-based hate speech/counter-narrative pairs dataset on Islamophobia (CONAN) [34] has been employed too.

5.1. Final Ranking

Table 3 shows the results obtained by the participants for each of the four subtasks. The runs submitted by each participant are highlighted in green. However, when a team submits a run to Task A - Textual, the submission satisfies also Task A - contextual and Task B - XPoliticalHate requirements, therefore it is included in the final ranking. Likewise, when a team submits a run to Task A - contextual, the submission satisfies Task B - XPoliticalHate requirements too. The best results in Task A - Textual, Task A - contextual, and Task B - XPoliticalHate

Table 3
Participants' results for each task and subtask. Results are reported as averaged F_1 scores. Best results are in bold. The runs submitted by each team are highlighted in green, whereas the remaining ones indicate the tasks and subtasks in which the team implicitly participated in.

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

are achieved by the **odang4** team with $F_{1(avg)}$ = 0.912, employing the same model without taking advantage of contextual information nor using external data sources. Only **extremITA** and **LMU** (the latter exclusively participated to Task B - XPoliticalHate) reached $F_{1(avg)} >$ 0.9 with at least one of their runs.

extremITA and **LMU** are the only two teams that reached $F_{1(avg)} > 0.6$ in Task B - XReligiousHate. In particular, **extremITA** obtained $F_{1(avg)}$ = 0.6525, with a remarkable improvement with respect to other teams.

All participating systems showed an improvement over the baselines employed for the in-domain political hate speech detection tasks, whereas only two teams outperformed the baseline for Task B - XReligiousHate, proving the complexity of the cross-domain task (Section 5.1).

6. Discussion and Conclusion

Results show that the run #1 submitted by the **odang4** team achieves the best scores across all in-domain tasks. In particular, their approach combining prompting, the Ontology of Dangerous Speech, and the ALBERTo model proved particularly effective in the political domain. However, none of the participants seems to have found a way to effectively exploit contextual information yielding an improvement over textual-only models. This is in line with past studies showing the challenges of embedding contextual information in hate speech detection systems [35].

While the best performance for the in-domain task confirms the state-of-the-art results obtained in similar settings [36], we observe a significant drop in performance (around -0.30 F_1 score on average) for the out-of-domain task. Among the systems, **extremITA** shows a better generalisation capability and yields the best results in this setting. We hypothesise that this happens because their system was built to address all EVALITA challenges, and the only task-specific adaptation is the use of instructions for HaSpeeDe3. Overall, out-of-domain settings still challenge hate speech detection capabilities and still represent a research direction to investigate. Furthermore, approaches that tackle well in-domain hate do not seem to suit the out-of-domain setting, for which different strategies should be pursued.

Acknowledgments

This work has received financial support from the European Union's Horizon Europe research and innovation program under grant agreement No 101070190 (AI4Trust).

References

[1] CENSIS, 50° rapporto sulla situazione sociale del paese 2016, Franco Angeli, 2016.

[2] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, A. Flammini, Political polarization on Twitter, in: International AAAI Conference on Web and Social Media, ICWSM, Association for the Advancement of Artificial Intelligence, Palo Alto, CA, USA, 2011, pp. 89–96.

[3] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE access 6 (2018) 13825–13835.

[4] F. Celli, E. A. Stepanov, M. Poesio, G. Riccardi, Predicting brexit: Classifying agreement is better than sentiment and pollsters., in: PEOPLES@ COLING, 2016, pp. 110–118.

[5] N. Oliver, B. Lepri, H. Sterly, R. Lambiotte, S. Deletaille, M. De Nadai, E. Letouzé, A. A. Salah, R. Benjamins, C. Cattuto, et al., Mobile phone data for informing public health actions across the covid-19 pandemic life cycle, 2020.

[6] M. Caprolu, A. Sadighian, R. Di Pietro, Characterizing the 2022 russo-ukrainian conflict through the lenses of aspect-based sentiment analysis: Dataset, methodology, and preliminary findings, 2022. URL: https://arxiv.org/abs/2208.04903. doi:10.48550/ARXIV.2208.04903.

[7] E. Elejalde, L. Ferres, E. Herder, The nature of real and perceived bias in chilean media, in: Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT, Association for Computing Machinery, New York, NY, USA, 2017, pp. 95–104. URL: http://doi.acm.org/10.1145/3078714.3078724. doi:10.1145/3078714.3078724.

[8] Y. Theocharis, W. Lowe, Does Facebook increase political participation? Evidence from a field experiment, Information, Communication & Society 19 (2016) 1465–1486.

[9] O. Ștefăniță, D.-M. Buf, Hate speech in social media and its effects on the lgbt community: A review of the current research, Romanian Journal of Communication and Public Relations 23 (2021).

[10] E. Fersini, D. Nozza, P. Rosso, Overview of the evalita 2018 task on automatic misogyny identification (ami), in: Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop, EVALITA 2018, volume 2263, CEUR, 2018.

[11] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, C. Bosco, Hate speech annotation: Analysis of an italian twitter corpus, in: 4th Italian Conference on Computational Linguistics, CLiC-it 2017, volume 2006, CEUR-WS, 2017, pp. 1–6.

[12] P. Badjatiya, S. Gupta, M. Gupta, V. Varma, Deep learning for hate speech detection in tweets, in: Proceedings of the 26th International Conference on World Wide Web Companion, 2017, pp. 759–760.

[13] F. Celli, G. Riccardi, A. Ghosh, Corea: Italian news corpus with emotions and agreement., in: Proceedings of CLIC-it 2014, 2014, pp. 98–102.

[14] M. Lai, M. Tambuscio, V. Patti, G. Ruffo, P. Rosso, Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter, Data & Knowledge Engineering 124 (2019) 101738. URL: https://www.sciencedirect.com/science/article/pii/S0169023X19300187. doi:https://doi.org/10.1016/j.datak.2019.101738.

[15] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources & Evaluation 55 (2021) 477–523.

[16] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hateminers: detecting hate speech against women, arXiv preprint arXiv:1812.06700 (2018).

[17] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Inf. Process. Manag. 57 (2020) 102360. URL: https://doi.org/10.1016/j.ipm.2020.102360. doi:10.1016/j.ipm.2020.102360.

[18] I. Guellil, A. Adeel, F. Azouaou, S. Chennoufi, H. Maafi, T. Hamitouche, Detecting hate speech against politicians in arabic community on social media, International Journal of Web Information Systems (2020).

[19] S. Jaki, T. De Smedt, Right-wing german hate speech on twitter: Analysis and automatic detection, arXiv preprint arXiv:1910.07518 (2019).

[20] A. Ramponi, B. Testa, S. Tonelli, E. Jezek, Addressing religious hate online: from taxonomy creation to automated detection, PeerJ Computer Science 8 (2022) e1128. URL: https://doi.org/10.7717/peerj-cs.1128. doi:https://doi.org/10.7717/peerj-cs.1128.

[21] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[22] A. Duzha, C. Casadei, M. Tosi, F. Celli, Hate versus politics: detection of hate against policy makers in italian tweets, SN Social Sciences 1 (2021) 1–15.

[23] F. Giglietto, N. Righetti, G. Marino, L. Rossi, Multiparty media partisanship attention score. estimating partisan attention of news media sources using twitter data in the lead-up to 2018 italian election, Comunicazione politica 20 (2019) 85–108.

[24] F. Celli, M. Lai, A. Duzha, C. Bosco, V. Patti, Policycorpus xl: An italian corpus for the detection of hate speech against politics, in: Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021), volume 3033 of *CEUR Workshop Proceedings*, CEUR-WS.org, Aachen, Germany, 2022. URL: http://ceur-ws.org/Vol-3033/paper38.pdf.

[25] M. Wiegand, J. Ruppenhofer, T. Kleinbauer, Detection of Abusive Language: the Problem of Biased Datasets, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 602–608. URL:

https://aclanthology.org/N19-1060. doi:10.18653/v1/N19-1060.

[26] A. Ramponi, S. Tonelli, Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 3027–3040. URL: https://aclanthology.org/2022.naacl-main.221. doi:10.18653/v1/2022.naacl-main.221.

[27] M. Sanguinetti, G. Comandini, E. Di Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Overview of the evalita 2020 second hate speech detection task (haspeede 2), in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR.org, Online, 2020.

[28] L. Grotti, P. Quick, Berticelli at haspeede3: Fine-tuning and cross-validating large language models for hate speech detection, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[29] I. Siragusa, R. Pirrone, Chilab at evalita 2023: Overview of the taks a textual, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[30] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at evalita 2023: Multi-task sustainable scaling to large language models at its extreme, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[31] V. Hangya, A. Fraserl, Lmu at haspeede3: Multi-dataset training for cross-domain hate speech detection, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[32] C. Di Bonaventura, A. Muti, M. A. Stranisci, B. McGillivray, A. Meroño-Peñuela, O-dang4 at hodi and haspeede3: A knowledge-enhanced approach to homotransphobia and hate speech detection in italian, EVALITA 2023 Eigth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (2023) –.

[33] M. A. Stranisci, S. Frenda, M. Lai, O. Araque, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, O-dang! the ontology of dangerous speech messages, in: Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data, European Language Resources Association, Marseille, France, 2022, pp. 2–8. URL: https://aclanthology.org/2022.salld-1.2.

[34] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, 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, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2819–2829. URL: https://aclanthology.org/P19-1271. doi:10.18653/v1/P19-1271.

[35] S. Menini, A. P. Aprosio, S. Tonelli, Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection, arXiv preprint arXiv:2103.14916 (2021).

[36] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, c. Çöltekin, SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020), in: Proceedings of SemEval, 2020.