O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian

Chiara Di Bonaventura[1], Arianna Muti[2] and Marco Antonio Stranisci[3]

[1]King's College London, United Kingdom
[2]University of Bologna, Italy
[3]University of Turin, Italy

Abstract
This paper describes our methods implemented during the EVALITA 2023 campaign for homotransphobia (HODI task) and hate speech detection (HaSpeeDe3 task) in Italian. We present three knowledge-enhanced approaches, namely via triple verbalisation, via prompting and via a majority vote, and we compare them to the AlBERTo baseline. These systems leverage the knowledge graph O-Dang, which contains information about named entities in Italian dangerous speech. Our knowledge-enhanced systems outperformed all the competition's baselines. Our best submissions achieved the macro-F1 score of 0.912 for HaSpeeDe3 and 0.795 for HODI, reaching the 1st and 3rd place, respectively. These results were achieved by using our baseline for HODI, and a majority voting approach for HaSpeeDe3.[1]

Keywords
hate speech, knowledge graph, entity linking, data augmentation, prompting

***Warning**: This paper contains examples of potentially offensive content.*[1]

1. Introduction

Technological progress and increasing online communication have made it necessary to create automatic tools for the detection of online abusive language to protect users. Indeed, online abusive language not only has increased over the past years, but also has effects that usually expand beyond the online context[2]. Usually, most of the targeted victims belong to minority groups because of their gender identity, sexual orientation, political and religious affiliation, *inter alia*. Therefore, their protection is of the utmost importance. Gender- and sexual-based violence manifests in social networks every time the abusive language harms LGBTQIA+ individuals directly or indirectly with homotransphobic discourse [2]. Political- and religious-based hate speech instead discriminates people based on their beliefs, affiliations, or ideologies.

Within the NLP community, many recent works propose solutions to automatically classify misogynous and sexist content [3, 4, 5], homotransphobic content [6, 7], and religious hate speech [8].

Evaluation campaigns have sped up the development of innovative approaches for shared tasks. One example is the Homotransphobia Detection in Italian (HODI) shared task [9], which was introduced during EVALITA 2023 [10]. The task comprises two subtasks. The first subtask, Subtask A - Homotransphobia detection, focuses on automatically identifying whether Italian online posts contain homotransphobic content or not. Subtask B - Explainability, aims at extracting the rationales of the classification model trained for Subtask A.

Another example is the shared task HaSpeeDe3 [11], presented during EVALITA 2023 to boost research on political and religious hate speech in Italian tweets. The task counts two subtasks. The goal of the first subtask, Subtask A - Political Hate Speech Detection, is to determine whether the message contains political hate speech or not. The problem is further divided in two distinct approaches, namely textual and contextual. In the former, participants can only use the provided textual content of the tweets, whereas in the latter, they can employ additional contextual information (e.g., metadata of the tweet and author, friends, etc.). Subtask B - Cross-domain Hate Speech Detection, instead, proposes a binary hate speech detection on test data belonging to different hate domains. Precisely, two settings are evaluated: XPolitical-Hate, where the participants can use external data from any kind of other domains, and XReligiousHate, where submissions are tested on tweets from the religious hate domain.

In this paper, we describe our proposed approach to ad-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
chiara.di bonaventura@kcl.ac.uk (C. D. Bonaventura); iarianna.muti2@unibo.it (A. Muti); marcoantonio.stranisci@unito.it (M. A. Stranisci)

[1]Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [1]

[2]https://www.ohchr.org/en/stories/2021/03/report-online-hate-increasing-against-minorities-says-expert

dress the HODI and HaSpeeDe3 shared tasks. We propose a knowledge-enhanced approach on top of the AlBERTo baseline that leverages external knowledge (System 1), internal knowledge (System 2) or both (System 3). Our results suggest that knowledge-enhancement can improve abusive language detection depending on the hate domain under study.

The rest of the paper is structured as follows. Section 2 describes the training datasets provided by the tasks' organisers whereas Section 3 describes our proposed systems. Section 4 summarises the experiments performed and discusses the results. Section 5 includes related work in the field. Section 6 draws some conclusions and discusses further possible research lines.

2. Data

To address the two tasks, datasets were provided by the tasks organisers. For HODI, we focus on Subtask A. 5,000 tweets were provided, manually labelled according to two classes, homotransphobic and not. Data are slightly skewed towards the negative class. For HaSpeeDe3, we focus on both Subtask A and B, where A aims to classify tweets between hateful and not, where hate is addressed towards politicians, while in Subtask B hate is addressed towards religious communities, in a cross-domain setting. PolicyCorpusXL (for Task A) contains 7,000 tweets about political debates, while ReligiousHate (for Task B) is composed by 3,000 tweets about the three main monotheistic religions, namely Christianity, Islam and Judaism.

3. Description of our Systems

In this work, we adopt a knowledge-enhanced approach to address the following subtasks: Subtask A for the HODI shared task, and Subtask A - textual and Subtask B - XReligiousHate for the HaSpeeDe3 shared task. However, the systems submitted for the Subtask A - textual satisfy the constraints for Subtask A - contextual and Subtask B - XPoliticalHate too.

Our intuition is to leverage knowledge about named entities in the training data and their association to online abusive language, which can provide auxiliary information to solve the tasks. Abusive language detection systems often fail to capture different nuances of abusive language because of the lack of contextual information [12] and the target-oriented nature of hate speech [13]. We propose knowledge-injection of relevant information into the system to help address this challenge. Firstly, we link training instances to the named entities associated with them. Then, we collect relevant information about named entities from the O-Dang knowledge graph and the Davinci OpenAI model, which are then injected into the AlBERTo baseline [14], a version of BERT for the Italian language trained on Twitter posts which includes emojis, links, hashtags, and mentions. AlBERTo was trained on 200M tweets randomly sampled from the TWITA corpus [15].

The following paragraphs describe the entity linking step, explain the knowledge-enhancement methods we implemented, and provide insights on its application to HODI and HaSpeeDe3 shared tasks.

3.1. Entity Linking

Our knowledge augmentation strategy relies on an entity linking pipeline based on a Knowledge Graph (KG) modelled on the Ontology of Dangerous Speech (O-Dang) [16]. The KG is a snapshot of a Wikidata [17] dump[3] where only entities of the type 'person' were retained, together with a set of 31 properties conveying relevant information (eg: 'member of political party' (P102), 'place of birth' (P19), spouse (P26)). As a result, we obtained 9,552,706 entities and 69,521,846 triples related to them.

After building the knowledge base, we implemented a pipeline for Entity Linking organized in three steps:

1. we identified all the entities of the type PERSON in the training sets of HaSpeeDe and HODI with Spacy[4];
2. we found all potential candidates of each detected entity, by using the Wikipedia APIs[5];
3. we generated three scores for each pair of the type <entity, candidate>: string similarity based on the Ratcliff/Obershelp pattern recognition [18], cosine similarity based on AlBERTo embeddings [14], the ranking of candidates returned from Wikipedia APIs. We retrieved 10 candidates and we only kept the most relevant result.

An example of such a pipleline is the following: Spacy identified 'Zorzi' as a named entity of the type person in the tweet: '@user_abcdefghij **Zorzi**, poveretto parla come una ch*cca isterica *@user_abcdefghij **Zorzi**, poor guy, he talks like a f*ggot*'. We queried Wikipedia APIs inputting the string 'Zorzi', obtaining the candidate 'Tommaso Zorzi', which was the first ranked of the list. We computed the Ratcliff/Obershelp pattern recognition to obtain a similarity score between 'Zorzi' and 'Tommaso Zorzi', which was 0.55 and we did the same with their embeddings (0.9). We averaged these scores to obtain a general score of 0.818. After a manual review of the scores, we decided to keep only linked entities with a score equal to or above 0.8.

[3] https://academictorrents.com/download/229cfeb2331ad43d4706efd435f6d78f40a3c438.torrent

[4] https://spacy.io/

[5] https://www.mediawiki.org/wiki/API:Search

The resulting number of entities identified in the two datasets and linked to our KG are 388 from HODI and 556 from HaSpeeDe3.

3.2. Knowledge-enhancement

We compare the AlBERTo baseline to the following three knowledge-enhanced systems:

System 1: enhanced-AlBERTo with triple verbalisation. For each named entity in the input data, we create a verbalised description using the information retrieved from O-Dang [16] through entity linking. Then, this description is concatenated at the end of the input instance, and passed to AlBERTo for the classification. We generate up to three different verbalised templates per O-Dang property to account for linguistic variation. For instance, the property 'P108' which stands for job location is translated into the following templates: (i) 'lavora presso' *(works at)*, (ii) 'svolge la sua professione presso' *(carries out their profession at)*, or (iii) 'svolge il suo lavoro presso' *(carries out their job at)*. When creating these templates, we ensure gender neutrality to avoid any possible source of bias. Then, for each input naming an O-Dang entity, the system randomly picks one template per property to create the verbalisation of the O-Dang triples whose heads refer to the named entity. As a result, the verbalised descriptions for the same named entity tend to be all linguistically different. Further, we fine-tune Italian BERT[6] on ItaCoLa [19] to check the linguistic acceptability of the descriptions. All descriptions passed the test since they were predicted as linguistically acceptable according to the model.[7] Table 1 shows statistics of the O-Dang triples and their verbalisation: the average number of triples, the max number of triples, and the average length of the verbalised triples of an O-Dang entity.

	HaSpeeDe3	HODI
avg #triples	9.9	9.7
max #triples	24	24
avg length verbalised triples	43	40

Table 1
Statistics of O-Dang triples and their verbalisation

System 2: enhanced-AlBERTo with prompting Davinci model. For the second system, we experiment with prompting large language models to leverage the knowlegde about abusive language they have acquired during pretraining. We try the following prompts, among others:

- **P1:** quanto è probabile che ENTITY scriva un tweet offensivo? Se sì, perchè? *(How likely is ENTITY to write an offensive tweet? If so why?)*
- **P2:** quanto è probabile che ENTITY sia associato ad un tweet offensivo? Se sì, perchè? *(How likely is ENTITY to be associated to an offensive tweet? If so why?)*
- **P3:** quanto è probabile che ENTITY sia vittima di un tweet offensivo? Se sì, perchè? *(How likely is ENTITY to be victim of an offensive tweet? If so why?)*

Among them, only prompt **P3** returned relevant results. Indeed, direct questions such as **P1** and **P2** triggered the language model into answering with a non-response — i.e., 'Come modello di intelligenza artificiale, non ho accesso a informazioni specifiche sulle associazioni di ENTITY con tweet offensivi' *(As an artificial intelligence model, I do not have access to specific information about ENTITY's associations with offensive tweets)*. On the other hand, the keyword 'victim' does not trigger the model, which is thus able to return reasonable answers given its pretraining knowledge. Further, we compared popular large language models in order to choose the best one. Table 2 and Table 3 show the difference in returned answers between Davinci and Curie OpenAI models, respectively. Davinci performs very well, whereas Curie – which according to the OpenAI website[8] performs almost equally well – performs poorly with respect to knowledge about offensive content.

Similar to System 1, Davinci's knowledge is injected into AlBERTo by concatenating the returned answer to prompt P3 at the end of the input instance, where ENTITY refers to the named entity in the instance. If the instance names multiple entities, we concatenate all their descriptions.

System 3: Majority vote. The third system takes a majority vote between the AlBERTo baseline, the enhanced-AlBERTo with triple verbalisation and the enhanced-AlBERTo with prompting Davinci. The intuition behind is to leverage both external knowledge provided by O-Dang triples and internal knowledge provided by Davinci model.

3.3. Approach to HODI Shared Task

Subtask A. In this task, we compare AlBERTo baseline to enhanced-AlBERTo with triple verbalisation (System 1) and enhanced-AlBERTo with prompting Davinci model (System 2).

[6] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[7] The fine-tuned BERT we used for the inference on the verbalised descriptions achieved a total MCC score of +0.57 on the ItaCoLa dataset.

[8] https://platform.openai.com/docs/models/gpt-3

	Davinci	Curie
OpenAI descriptions	Most capable GPT-3 model. Can do any task the other models can do, often with higher quality.	Very capable, but faster and lower cost than Davinci.
Prompt P3 with ENTITY = Matteo Salvini	Il leader della Lega Nord, Matteo Salvini, è probabilmente vittima di commenti online offensivi a causa della sua politica anti-immigrazione. Salvini ha preso una posizione molto dura contro l'immigrazione clandestina e ha fatto campagne per limitare l'ingresso di immigrati in Italia. La sua politica ha attirato molte critiche da parte di coloro che sostengono che l'immigrazione è un diritto umano fondamentale. Inoltre, Salvini è spesso criticato *The leader of Lega Nord, Matteo Salvini, is probably the victim of offensive online comments because of his anti-immigration policy. Salvini took a very tough stance against illegal immigration and campaigned to limit the entry of immigrants into Italy. His policy has attracted much criticism from those who argue that immigration is a fundamental human right. In addition, Salvini is often criticized*	Se è probabile che un commento online offensivo possa costare una condanna penale, perché non si cerca di intervenire? Perché non si fa niente? Perché non si fa niente? Perché non si fa niente? Perché non si fa niente? *If an offensive online comment is likely to cost a criminal conviction, why is there no attempt to intervene? Why is nothing done? Why is nothing done? Why is nothing done? Why is nothing done?*

Table 2
Comparison between Davinci and Curie OpenAI models on how they reply to the question 'How likely is Matteo Salvini to be victim of an offensive tweet? If so why?' (P3)

	Davinci	Curie
Prompt P3 with ENTITY = Matteo Renzi	Nel corso degli ultimi anni, Matteo Renzi è stato vittima di diversi commenti online offensivi. In particolare, nel 2016, è stato preso di mira da alcuni utenti che hanno criticato il suo modo di gestire la crisi economica in Italia. *Over the past few years, Matteo Renzi has been the victim of several offensive online comments. In particular, in 2016, he was targeted by some users who criticized his way of managing the economic crisis in Italy.*	Il commento è stato dato da una persona molto influente, che è il presidente del partito, che ha dato una risposta che l'ha definita "inaccettabile". Il partito non è una scuola di formazione, ma una scuola di vita, e quindi la risposta deve essere meno formale, e più educata. Perché questa persona è il presidente del partito? C'è una persona che rappresenta il partito, che è una persona che rappresenta la linea *The comment was given by a very influential person, who is the president of the party, who gave a response that called it "unacceptable". The party is not a training school, but a school of life, and therefore the answer must be less formal, and more educated. Why is this person the party president? There is a person who represents the party, who is a person who represents the political line*

Table 3
Comparison between Davinci and Curie OpenAI models on how they reply to the question 'How likely is Matteo Renzi to be victim of an offensive tweet? If so why?' (P3)

3.4. Approach to HaSpeeDe3 Shared Task

Subtask A - textual. For hate speech detection in the political domain, we submit our System 2 (baseline enhanced with Davinci) and the majority vote (System 3).

Subtask B - XReligiousHate. Similarly, we compare System 2 and System 3. Since subtask B explores cross-

domain hate speech detection by focusing on the religious domain, we further augment the training dataset with CONAN [8], which is a multilingual expert-based hate speech/counter-narrative pairs dataset on Islamophobia. More precisely, we use only Italian hate speech content for augmentation.

4. Experimental Setup and Results

Setting For the baseline, we fine-tune AlBERTo to our downstream tasks. We perform a minimum parameter selection tuning on the validation set (10% of the training set). We selected the highest performing learning rate $\in [1e{-}5, 2e{-}5, 1e{-}2]$; batch size $\in [4, 8, 16, 32]$; epochs in range $[1 - 10]$. The best configuration for both models is: lr = 1e−5, batch size = 16, epochs = 4. In order to tune the network, we used the AdamW optimizer. As for the pre-processing, we used the pretrained AlBERTo tokenizer for text tokenization, and then we encoded the data. We set the maximum length to 256 characters for the baseline, and 512 for instanced enriched with verbalisation or prompting.

For System 2, we use *text-davinci-002* with *temperature=0.5*. To avoid too long answers, particularly for multiple named entities, we set the max number of returned sentences to *n=5* with *max_tokens=150*.

Results Tables 4 and 5 show the results on the development and test set of the HODI and HaSpeeDe3 shared tasks. While for HODI we submitted all models, for HaSpeeDe3 only the models that performed the best in the development set were subsequently submitted. For HaSpeeDe3 - Task B we could not produce results on the development set because the religious data were only released during the testing phase.

Setting	Dev	Test
AlBERTo (run 1)	0.83	0.795
Davinci (run 2)	0.87	0.792
Verbalisation (run 3)	0.82	0.780
Majority Vote (not submitted)	-	0.80
baseline	-	0.669
top 1	-	0.810

Table 4
Experiments and macro-averaged F_1 scores on the development and test set for HODI. We report also the scores of the baseline produced by the organisers and the top-performing team.

In both tasks, our models struggle the most with the identification of the positive class, which is the least represented in the training data. For HODI, enhancing the baseline with the text generated from Davinci helps us gain 0.04 points with respect to the baseline in the development set, but we observe a slight drop in the test set, leading our baseline to be our best submitted system. In both the development and test set, enhancing the baseline with the verbalisation cause a drop of 0.01 point with respect to the baseline. Majority voting does not prove to be effective in this case. For HaSpeeDe3, we obtain the best result by enhancing our baseline with Davinci, gaining 0.01 point. Combining both the verbalisation and Davinci leads to a drop of 0.01 point instead, while adding only the verbalisation does not affect the final score. However, in the test set we obtain the best result with majority vote, gaining two units over the best-performing model, i.e., Davinci.

	Task A		Task B
Setting	dev	test	test
AlBERTo	0.91	0.91	0.51
Davinci (run 2)	0.92	0.89	0.48
Verbalisation	0.91	0.91	0.51
Davinci + Verbalisation	0.90	0.89	0.47
Majority Vote (run 1)	-	0.91	0.52

Table 5
Experiments and macro-averaged F_1 scores on the development and test set for Subtasks A and B of HaSpeeDe3.

Error Analysis We performed an analysis of the classification errors over our submitted systems. First, with the help of NLTK, we retrieved the most frequent words in misclassified instances. In both tasks, enhancing the models with a knowledge-base, either through Davinci or the verbalisation, results in less misclassified instances containing name entities, proving the efficacy of our approach, at the expenses of other instances that do not contain name entities hence are not enhanced, which do not get identified correctly. For what concerns HaSpeeDe3, most false positives comes from sentences showing a negative sentiment towards racist statements, like in the following example: *Se la SeaWatch fosse piena di gattini sareste già partiti con i pedalò per salvarli. Mi disgustate.*[9] As for false negatives, we observed a pattern of misclassification when the target of the statement is implicit, like in the following example: *Quelli che si lamentano della puzza di urina per le strade di Roma sono gli stessi che dicono #Fateliscendere o #portiaperti ! Ma secondo loro chi c*zzo è che piscia per strada a tutte le ore, che vagano ubriachi o senza meta? Gli alieni?.*[10] In this case, there is an implicit stereotyped racist statement, i.e., those who pee on the streets are all immigrants.

[9] If SeaWatch was full of kittens you would have already left with paddleboats to rescue them. You disgust me.

[10] Those who complain about the stench of urine on the streets of Rome are the same who say #Letthemin or #opentheports ! But according to them who the f*ck is pissing in the streets at all hours, wandering drunk or aimlessly? The aliens?

5. Related Work

Recent work has shown the efficiency of knowledge-enhancement of NLP models in many downstream tasks, such as sentiment classification [20], word sense disambiguation [21], and semantic change detection [22].

Sharifirad et al. [23] is one of the first attempts to leverage external world knowledge for abusive language detection, improving performance on sexist tweet classification. They use ConceptNet [24, 25] and Wikidata [17] to augment the original data by concatenating additional information about concepts and their descriptions. Similarly, Lin [26] uses an entity linking approach to link entities mentioned in tweets to their Wikipedia descriptions in order to leverage world knowledge for hate speech detection. The injection of external world knowledge is a promising avenue for explicit hate speech detection, leading to an improvement of 10% for precision, recall and F1-score. We build upon these works, and leverage both external and internal knowledge about abusive language to explore their impact on different hate domains. To the best of our knowledge, we are the first to apply internal knowledge of large language models through prompt-engineering to the hate speech detection task, while most of the works focus on natural language understanding, question answering or text completion [27].

6. Conclusions

We present a knowledge-enhanced classification solution for identifying homotransphobic and hate speech content in Italian online posts. Our first system uses external knowledge injection via O-Dang triple verbalisation to enhance the AlBERTo model, whereas our second system exploits Davinci's internal knowledge about abusive language to enhance AlBERTo model. Lastly, a majority vote among AlBERTo baseline, System 1 and System 2 is adopted to improve classification particularly on uncertain predictions. We evaluate our approach in the Homotransphobia Detection in Italian (HODI) and Hate Speech Detection (HaSpeeDe3) Shared Tasks of the EVALITA 2023 campaign. Our results show that knowledge-enhancement can improve the classification, especially of sentences containing name entities in the political hate domain. The resulting approach can be expanded on multiple knowledge sources, knowledge injection methods and tasks.

Acknowledgments

The work of Chiara Di Bonaventura was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). CDB would like to thank her supervisors, Albert Meroño-Peñuela and Barbara McGillivray, for their helpful comments and mentorship.

References

[1] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[2] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021).

[3] E. Fersini, D. Nozza, G. Boifava, Profiling Italian misogynist: An empirical study, in: Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language, European Language Resources Association (ELRA), Marseille, France, 2020, pp. 9–13. URL: https://aclanthology.org/2020.restup-1.3.

[4] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360.

[5] H. R. Kirk, W. Yin, B. Vidgen, P. Röttger, Semeval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, 2023. URL: http://arxiv.org/abs/2303.04222. doi:10.48550/arXiv.2303.04222.

[6] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on Twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 16–24. URL: https://aclanthology.org/2023.c3nlp-1.3.

[7] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitaleer, P. K. Kumaresan, R. Ponnusamy, Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2022.

[8] 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.

[9] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, 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.

[10] 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.

[11] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task, 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.

[12] B. Vidgen, D. Nguyen, H. Margetts, P. Rossini, R. Tromble, Introducing CAD: the contextual abuse dataset, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 2289–2303. URL: https://aclanthology.org/2021.naacl-main.182. doi:10.18653/v1/2021.naacl-main.182.

[13] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114. doi:10.18653/v1/2021.acl-short.114.

[14] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets, in: Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019), volume 2481, CEUR, 2019. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

[15] V. Basile, M. Lai, M. Sanguinetti, Long-term social media data collection at the university of turin, in: Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Turin, Italy, 2018.

[16] 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.

[17] D. Vrandečić, M. Krötzsch, Wikidata: A free collaborative knowledgebase, Commun. ACM 57 (2014) 78–85. URL: https://doi.org/10.1145/2629489. doi:10.1145/2629489.

[18] J. Ratcliff, D. Metzener, Ratcliff-obershelp pattern recognition, Dictionary of Algorithms and Data Structures (1998).

[19] D. Trotta, R. Guarasci, E. Leonardelli, S. Tonelli, Monolingual and cross-lingual acceptability judgments with the Italian CoLA corpus, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2929–2940. URL: https://aclanthology.org/2021.findings-emnlp.250. doi:10.18653/v1/2021.findings-emnlp.250.

[20] P. Ke, H. Ji, S. Liu, X. Zhu, M. Huang, SentiLARE: Sentiment-aware language representation learning with linguistic knowledge, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6975–6988. URL: https://aclanthology.org/2020.emnlp-main.567. doi:10.18653/v1/2020.emnlp-main.567.

[21] J. Zhou, Z. Zhang, H. Zhao, S. Zhang, LIMIT-BERT : Linguistics informed multi-task BERT, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 4450–4461. URL: https://aclanthology.org/2020.findings-emnlp.399. doi:10.18653/v1/2020.findings-emnlp.399.

[22] B. McGillivray, M. Alahapperuma, J. Cook, C. Di Bonaventura, A. Meroño-Peñuela, G. Tyson, S. Wilson, Leveraging time-dependent lexical features for offensive language detection, in: Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 39–54. URL: https://aclanthology.org/2022.evonlp-1.7.

[23] S. Sharifirad, B. Jafarpour, S. Matwin, Boosting text classification performance on sexist tweets

by text augmentation and text generation using a combination of knowledge graphs, in: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 107–114. URL: https://aclanthology.org/W18-5114. doi:10.18653/v1/W18-5114.

[24] H. Liu, P. Singh, Conceptnet—a practical commonsense reasoning tool-kit, BT technology journal 22 (2004) 211–226.

[25] R. Speer, J. Chin, C. Havasi, Conceptnet 5.5: An open multilingual graph of general knowledge, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, AAAI Press, 2017, p. 4444–4451.

[26] J. Lin, Leveraging world knowledge in implicit hate speech detection, in: Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 31–39. URL: https://aclanthology.org/2022.nlp4pi-1.4.

[27] R. Brate, M.-H. Dang, F. Hoppe, Y. He, A. Meroño-Peñuela, V. Sadashivaiah, Improving language model predictions via prompts enriched with knowledge graphs, in: Workshop on Deep Learning for Knowledge Graphs (DL4KG@ ISWC2022), 2022.