An Exploration of Zero-Shot Natural Language Inference-Based Hate Speech Detection

Nerses Yuzbashyan
University of Antwerp
Belgium
nerses.yuzbashyan@uantwerpen.be

Nikolay Banar
University of Antwerp
Belgium
nicolae.banari@uantwerpen.be

Ilia Markov
Vrije Universiteit Amsterdam
The Netherlands
i.markov@vu.nl

Walter Daelemans
University of Antwerp
Belgium
walter.daelemans@uantwerpen.be

Abstract

Conventional techniques for detecting online hate speech rely on the availability of a sufficient number of annotated instances, which can be costly and time consuming. For this reason, zero-shot or few-shot detection can offer an attractive alternative. In this paper, we explore a zero-shot detection approach based on natural language inference (NLI) models. The performance of the models in this approach depends heavily on the choice of a hypothesis, which represents a statement that is evaluated with a given sentence to determine the logical relationship between them. Our goal is to determine which factors affect the quality of detection. We conducted a set of experiments with three NLI models and four hate speech datasets. We demonstrate that a zero-shot NLI-based approach is competitive with approaches that require supervised learning, yet they are highly sensitive to the choice of hypothesis. In addition, our experiments indicate that the results for a set of hypotheses on different model-data pairs are positively correlated, and that the correlation is higher for different datasets when using the same model than it is for different models when using the same dataset. These results suggest that if we find a hypothesis that works well for a specific model and domain or for a specific type of hate speech, we can use that hypothesis with the same model also within a different domain. While another model might require different suitable hypotheses in order to demonstrate high performance.

1 Introduction

The growing use of social media platforms that allow users to remain anonymous during online discussions has led to an increase in the amount of hateful content online. This has posed a challenge in detecting hate speech for government organizations, social media platforms, and the research community. Effective hate speech detection models that are robust and reliable can provide valuable insights to moderators in their efforts to combat the prevalence of hate speech in online discussions, as well as encourage productive online discourse (Halevy et al., 2022). In this paper, we use the term *hate speech* as an umbrella term for different types of insulting content, such as offensive language, abusive language, and other types of harmful content.

Since supervised learning methods are associated with difficulties such as the need for large computing power and extensive amounts of labeled data, zero-shot learning using pre-trained language models can be an attractive alternative. Zero-shot detection is a technique that allows a model to classify texts based on their content, even if the model has not been trained for that particular task (Larochelle et al., 2008). The main advantage of the zero-shot approach is its versatility. A model pre-trained on general data can be used to detect hate speech across multiple platforms (Facebook, Twitter, etc.) and domains (different targets and types of hate speech) without having to retrain it. This reduces training costs and allows for greater flexibility in responding to changes in social media platforms, user behavior, and types of hate speech. However, since the model is not specifically trained for the task of detecting hate speech, the approach might demonstrate inferior results to the supervised models. In this paper, we investigate whether an NLI-based zero-shot approach is competitive to supervised learning

LT-EDI 2023 – Third Workshop on Language Technology for Equality, Diversity and Inclusion,
pages 1–9, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-084-7_001

methods and explore its robustness for different models, datasets and targets of hate speech.

2 Related Work

Hate speech detection. Early approaches for hate speech detection were based on manual feature engineering (Burnap and Williams, 2015; Davidson et al., 2017; Waseem and Hovy, 2016). The majority of the current methods for detecting hate speech rely on one of two techniques: training machine learning models from scratch or fine-tuning pre-trained language models (Jahan and Oussalah, 2021; Markov et al., 2021; Uzan and HaCohen-Kerner, 2021; Banerjee et al., 2021; Nghiem and Morstatter, 2021; Markov et al., 2022). All of these methods require extensive amounts of labeled data, which is not consistently accessible for some languages (Poletto et al., 2021), and is extremely expensive to be annotated manually. Under these circumstances, zero-shot or few-shot detection may be an appealing option.

Zero-shot in hate speech detection. Ke-Li et al. (2021) used GPT-3 (Brown et al., 2020) to identify sexist and racist text passages with zero-shot, one-shot, and few-shot learning. They achieved an accuracy as high as 85% for few-shot learning and assumed that large language models with further development could eventually be used to detect hate speech. Yin et al. (2019) demonstrated that text classification tasks can be approached as natural language inference (NLI), resulting in high accuracy for zero-shot classification. Based on Yin et al. (2019), Goldzycher and Schneider (2022) developed strategies that aim at improving NLI-based zero-shot hate speech detection systems and showed that such approaches are able to outperform fine-tuned language models (acc. 79.4 for NLI zero-shot for the best performing hypothesis against acc. 76.6 for a fine-tuned model). However, NLI approaches require a *hypothesis* - a statement that is evaluated for its logical relationship with the target sentence. The performance of such approaches largely depends on the chosen hypothesis, and evaluation of the quality of each hypothesis may not be feasible. Another uncertainty lies in the formulation of supporting hypotheses. An inadequately formulated supporting hypothesis can have a detrimental impact on the model performance (Goldzycher and Schneider, 2022). Our goal is to evaluate an NLI-based approach for various models and datasets, in order i) to find out how the choice of a hypothesis affects the quality of the model, ii) to find out how the results change for a given hypothesis when the model or domain is changed, and iii) to determine which factors affect the accuracy of NLI-based zero-shot classification.

3 Method and Models

In order to determine whether an input text contains hate speech, we need a hypothesis that expresses that claim. In NLI tasks, the hypothesis is a statement (e.g., "This text is racist") that needs to be either supported or contradicted by a given premise. It is typically formulated as a sentence that makes a claim or draws a conclusion based on the information presented in the premise. All experiments were conducted in a standard setup for NLI-based zero-shot classification. We feed the hypothesis with an example to a pretrained NLI model and get the probability of entailment for the target sentence and hypothesis. We ignore the logits for *neutral* and perform a softmax over the logits of *contradiction* and *entailment*. We use the coarse-grained (binary) hate speech classes: hate speech versus non-hate speech. If the probability for entailment is equal or higher than 0.5 we consider that it is hate speech. We report the results in terms of F1-score (macro-averaged).

We conduct our experiments using the following well-established models, which are available via the Huggingface transformers library (Wolf et al., 2020):

- flan-t5-large (Chung et al., 2022): T5-large model (Raffel et al., 2020) was fine-tuned on a collection of NLI datasets. The full list of datasets and fine-tuning process is described in (Chung et al., 2022).
- bart-large-mnli (Williams et al., 2017): BART-large model (Lewis et al., 2019) was fine-tuned on the Multi-Genre Natural Language Inference dataset (MNLI (Williams et al., 2017)).
- XLM-RoBERTa-large-XNLI-ANLI: RoBERTa-large model (Liu et al., 2019) is fine-tuned on the ANLI (Nie et al., 2019) and XNLI (Conneau et al., 2018) datasets.

Dataset	Test set size	Classes	% (#)
FRENK (Ljubešić et al., 2019)	2,095	hate speech	35.5 (744)
		not hate speech	64.5 (1,351)
HateCheck (Röttger et al., 2020)	3,728	hate speech	68.8 (2,563)
		not hate speech	31.2 (1,165)
CAD (Vidgen et al., 2021)	5,307	hate speech	16.9 (899)
		not hate speech	83.1 (4,408)
OLID (Zampieri et al., 2019)	860	hate speech	27.9 (240)
		not hate speech	72.1 (620)

Table 1: Statistics of the datasets used.

4 Datasets

We evaluated the models described in Section 3 on four datasets constructed from different online platforms, covering different topics, types and targets of hate speech. We used the binary hate speech classes: hate speech versus non-hate speech. The statistics of the datasets used are shown in Table 1.

FRENK (Ljubešić et al., 2019). The FRENK dataset includes comments from Facebook on LGBT and migrants topics in English. The dataset was manually annotated for fine-grained types of socially unacceptable discourse (e.g., violence, offensiveness, threat). Messages were assigned to a particular class if at least four out of eight annotators agreed on the class. The test set consists of 2,095 examples.

HateCheck (Röttger et al., 2020) is an English, synthetic, evaluation-only dataset annotated for binary hate speech classification. For generating the test set, templates were prepared that contained one blank space to be filled with a discriminated group: women, gay people, transgender people, black people, Muslims, immigrants, and disabled people. The templates for non-hateful content share linguistic features with hateful expressions and could be mistaken for hate speech by a classifier. In total, the dataset consists of 3,728 examples.

CAD (Vidgen et al., 2021). The Contextual Abuse Dataset (CAD) consists of 25,000 annotated Reddit entries. All entries were first independently annotated by two annotators. Annotators worked through entire Reddit conversations, making annotations for each entry with full knowledge of the previous content in the thread. The test set consists of 5,307 examples.

OLID (Zampieri et al., 2019). The Offensive Language Identification Dataset (OLID) consists of 14,100 tweets in English, annotated through crowdsourcing. During annotation, each example was initially labeled by two annotators. In the case of disagreement, a third annotation was requested, and then a majority vote was taken. The test set consists of 860 entries.

5 Experiments and Results

In this section, we present the key findings and analysis derived from our research. We first report performance of the models on a general set of hypotheses. Then we reduce the number of hypotheses and use only those that describe a certain type or target of hate speech.

Evaluation of zero-shot detection. In the first series of experiments, we tested 111 manually designed hypotheses, describing hate speech, in order to determine how classification results vary depending on the chosen hypothesis, as well as which factors affect the results. Table 2 shows the comparison of a state-of-the-art supervised learning approach with the zero-shot approach. We report the results in terms of macro-average F1-score. We compare our zero-shot results with the supervised learning results reported in Markov and Daelemans (2021) for BERT, RoBERTa, SVM and a hard majority-voting ensemble of those three models applied to the FRENK and OLID datasets. In Markov and Daelemans (2021), models were tested in in-domain and cross-domain settings (trained on FRENK, tested on OLID and vice versa). Table 2 shows that the zero-shot NLI-based approach remains inferior to the supervised mod-

Method	Model	FRENK		OLID	
		In-domain	Cross-domain	In-domain	Cross-domain
SL	BERT	78.30	69.40	82.20	72.10
	RoBERTa	78.50	69.20	80.00	72.40
	SVM	77.00	67.70	78.30	69.00
	Ensemble	79.70	69.70	83.20	73.60
ZS	flan-t5-large	73.43		78.45	
	bart-large-mnli	72.39		73.30	
	xlm-roberta-large-xnli-anli	70.37		72.76	

Table 2: Comparison of supervised learning approach and zero-shot approach. We report macro-averaged F1-score. SL - supervised learning, ZS - zero-shot. For ZS, we report the result of the best hypothesis for each dataset-model pair. For SL, we report results from Markov and Daelemans (2021) for in-domain setting (models were trained and tested on the same dataset) and cross-domain setting (trained on FRENK, tested on OLID and vice versa).

els in the in-domain setting but surpasses it in the cross-domain setup. This indicates that a zero-shot method is more versatile and, in cases where there is not enough data for training for the current domain (e.g., in a case of "cold start"), a zero-shot approach could be preferred. However, as we can see from Table 3, even limited paraphrasing of the hypothesis can cause considerable variation in F1-score.

Hypothesis	F1	△F1
This text contains hateful content	71.80	
It contains hateful content	65.76	-6.04
This text contains xenophobic content	66.75	
This text is xenophobic	55.78	-10.97
It contains xenophobic content	52.98	-13.77
It is racist	69.14	
This text is racist	67.91	-1.23
This text contains racist content	66.49	-2.65
It contains racist content	63.12	-6.02

Table 3: Examples of variations in F1-score with minor paraphrasing of hypotheses for the FRENK dataset and flan-T5 model pair.

To investigate this variability more systematically, for each model-dataset pair, we built vectors whose elements are the F1-scores for the used hypotheses (in total 12 vectors of length 111, the hypotheses were sorted alphabetically) and calculated the correlation matrix for these 12 vectors (see Appendix A). One can see that the results for all model-dataset pairs are positively correlated, except for XLM-Roberta with the CAD dataset. The matrix Table 5 shows that, on average, the correlation for a particular model and different datasets is higher than the correlation for a dataset and different models.

Experiment with a small set of target hypotheses. In the second set of experiments, we used only the hypotheses that described a certain type of hate speech or certain target of hate speech (e.g., "This text is racist", "This text is homophobic", "This text is sexist", etc.), hence, we excluded "general" hypotheses (e.g. "This text is hateful", "This text contains hate speech", etc.). This experiment aimed to determine whether the performance of a particular hypothesis depends on which hate speech types are represented in a test dataset. In this case we expected that the dataset-related hypotheses will perform better, while another hypotheses will show a lower F1-score, and as a consequence there will be no correlation of results for different datasets.

However, we again observe that the results for different model-dataset pairs are positively correlated. Moreover, we see that the correlation of the results for different models when using the same dataset is lower on average than the correlation of the results for different datasets when using the same model (see Appendix B). From this, we can conclude that when choosing a hypothesis, it is more important to focus on what the model understands as hate speech rather than the type of hate speech covered in a particular dataset.

Experiment for test subsets covering a particular hate speech target. In order to verify our conclusion from the previous set of experiments, we split the FRENK test set into two subsets, each of which covers only one target of hate speech (LGBT or migrants). We observed that the hypotheses related to the topic of the test subset are

	FRENK LGBT		FRENK Migrants	
flan-t5	Top5 Hypothesis	F1	Top5 Hypothesis	F1
	This text is racist	67.47	This text is misandric	75.11
	This text is misogynistic	66.30	This text is racist	66.38
	This text is misandric	64.55	This text is hostile to migrants	61.59
	This text is hostile to lgbtq+ community	62.21	This text is hostile to immigrants	61.33
	This text is hostile to lgbt community	61.12	This text is xenophobic	55.63
bart	Top5 Hypothesis	F1	Top5 Hypothesis	F1
	This text is hostile to lgbt community	68.54	This text is hostile to migrants	64.72
	This text is hostile to woman	68.12	This text is hostile to immigrants	62.32
	This text is hostile to man	67.77	This text is xenophobic	56.08
	This text is hostile to lgbtq+ community	67.29	This text is hostile to woman	55.57
	This text is misogynistic	66.69	This text is misandric	54.49
roberta	Top5 Hypothesis	F1	Top5 Hypothesis	F1
	This text is xenophobic	67.71	This text is hostile to lgbtq+ community	68.09
	This text is sexist	67.69	This text is hostile to immigrants	66.85
	This text is woman-hatred	66.67	This text is misogynistic	66.50
	This text is racist	64.96	This text is xenophobic	66.03
	This text is man-hatred	64.39	This text is man-hatred	65.60

Table 4: Top 5 hypotheses in the experiment for test subsets per target of hate speech.

on average higher, though not always in the first position (see Table 4). The results confirm that on average the scores for the same model have a positive correlation (except for xlm-roberta). A positive correlation shows that even with hypotheses not related to the type of hate speech in the dataset, the model can still perform well. Additionally, it shows that it is important what the model understands by hate speech, although the topical focus of the dataset also affects the results.

6 Conclusion and Future Work

The first set of experiments showed that despite the fact that an NLI-based approach can compete with supervised methods, this approach is sensitive to the choice of hypothesis and even limited paraphrasing can change F1-scores substantially. Our experiments indicate that the results using particular (sets of) hypotheses for different model-data pairs are positively correlated, and that correlation for a particular model and different datasets is higher than the correlation for a dataset and different models. This suggests that if we find a hypothesis that works well for a specific model and dataset or a specific type of hate speech, we can use the same hypothesis for the same model but a different dataset. However, if the model is changed, it is better to search for an alternative hypothesis.

In future work, we plan to experiment with automatic hypothesis engineering. We want to answer the following questions: can we automatically find a better hypothesis than the initial one and will the hypothesis optimized for one data-model pair work well for other models, other domains and other data-model pairs.

References

Somnath Banerjee, Maulindu Sarkar, Nancy Agrawal, Punyajoy Saha, and Mithun Das. 2021. Exploring transformer based models to identify hate speech and offensive content in english and indo-aryan languages. *arXiv preprint arXiv:2111.13974*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Janis Goldzycher and Gerold Schneider. 2022. Hypothesis engineering for zero-shot hate speech detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 75–90, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Communications of the ACM*, 65(2):92–98.

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint arXiv:2106.00742*.

Chiu Ke-Li, Collins Annie, and Alexander Rohan. 2021. Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The frenk datasets of socially unacceptable discourse in slovene and english. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22*, pages 103–114. Springer.

Ilia Markov and Walter Daelemans. 2021. Improving cross-domain hate speech detection by reducing the false positive rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. Association for Computational Linguistics.

Ilia Markov, Ine Gevers, and Walter Daelemans. 2022. An ensemble approach for Dutch cross-domain hate speech detection. In *Proceedings of the 27th International Conference on Natural Language & Information Systems*, pages 3–15, Valencia, Spain. Springer.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Huy Nghiem and Fred Morstatter. 2021. "stop asian hate!": Refining detection of anti-asian hate speech during the covid-19 pandemic. *arXiv preprint arXiv:2112.02265*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Moshe Uzan and Yaakov HaCohen-Kerner. 2021. Detecting hate speech spreaders on twitter using lstm and bert in english and spanish. In *CLEF (Working Notes)*, pages 2178–2185.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. 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*, pages 2289–2303.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame,

Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Appendices

A Evaluation of Zero-Shot Detection. Correlation Matrix.

	t5 FR	t5 HC	t5 CAD	t5 OLID	bart FR	bart HC	bart CAD	bart OLID	xlm-rb FR	xlm-rb HC	xlm-rb CAD	xlm-rb OLID
t5 FR	1	0.76	0.82	0.83	0.3	0.43	-0.01	0.24	0.28	0.39	-0.14	0.21
t5 HC	0.76	1	0.84	0.8	0.19	0.5	0.13	0.19	0.33	0.34	0.01	0.3
t5 CAD	0.82	0.84	1	0.91	0.31	0.55	0.16	0.37	0.22	0.48	-0.2	0.21
t5 OLID	0.83	0.8	0.91	1	0.25	0.47	0.08	0.36	0.23	0.48	-0.21	0.27
bart FR	0.3	0.19	0.31	0.25	1	0.75	0.32	0.8	0	0.26	-0.27	-0.15
bart HC	0.43	0.5	0.55	0.47	0.75	1	0.1	0.72	0.09	0.38	-0.26	-0.04
bart CAD	-0.01	0.13	0.16	0.08	0.32	0.1	1	0.22	0.25	0.08	0.17	0.12
bart OLID	0.24	0.19	0.37	0.36	0.8	0.72	0.22	1	0.02	0.35	-0.3	0.02
xlm-rb FR	0.28	0.33	0.22	0.23	0	0.09	0.25	0.02	1	0.45	0.63	0.78
xlm-rb HC	0.39	0.34	0.48	0.48	0.26	0.38	0.08	0.35	0.45	1	-0.16	0.47
xlm-rb CAD	-0.14	0.01	-0.2	-0.21	-0.27	-0.26	0.17	-0.3	0.63	-0.16	1	0.55
xlm-rb OLID	0.21	0.3	0.21	0.27	-0.15	-0.04	0.12	0.02	0.78	0.47	0.55	1

Table 5: Correlation matrix for the experiment with hundred hypotheses. *FR* - FRENK, *HC* - HateCheck. We build the vectors of the results for every model-dataset combination. These vectors consist of F1-scores for the corresponding hypotheses. In total, there are 12 vectors, each of which with a length of 111. The hypotheses are sorted alphabetically, and the corresponding hypothesis vectors are used to compute the correlation matrix.

B Experiment with a Small Set of Target Hypotheses. Correlation Matrices.

FRENK	flan-t5	bart-large	xlm-roberta
flan-t5	1	0.53	0.34
bart-large	0.53	1	0.26
xlm-roberta	0.34	0.26	1

CAD	flan-t5	bart-large	xlm-roberta
flan-t5	1	0.61	-0.09
bart-large	0.61	1	-0.2
xlm-roberta	-0.09	-0.2	1

HateCheck	flan-t5	bart-large	xlm-roberta
flan-t5	1	0.33	0.09
bart-large	0.33	1	0.53
xlm-roberta	0.09	0.53	1

OLID	flan-t5	bart-large	xlm-roberta
flan-t5	1	0.45	0.21
bart-large	0.45	1	0.1
xlm-roberta	0.21	0.1	1

Table 6: Correlation of results for each dataset for different models in the experiment with a small set of target hypotheses.

flan-t5

	FRENK	HateCheck	CAD	OLID
FRENK	1	0.69	0.6	0.68
HateCheck	0.69	1	0.74	0.73
CAD	0.6	0.74	1	0.87
OLID	0.68	0.73	0.87	1

bart-large

	FRENK	HateCheck	CAD	OLID
FRENK	1	0.77	0.64	0.69
HateCheck	0.77	1	0.86	0.7
CAD	0.64	0.86	1	0.79
OLID	0.69	0.7	0.79	1

xlm-roberta

	FRENK	HateCheck	CAD	OLID
FRENK	1	0.54	0.44	0.68
HateCheck	0.54	1	-0.25	0.38
CAD	0.44	-0.25	1	0.44
OLID	0.68	0.38	0.44	1

Table 7: Correlation of results for each model for different datasets in experiment with small set of target hypotheses.