Analysis of Socially Unacceptable Discourse with Zero-shot Learning

Rayane Ghilene, Dimitra Niaouri, Michele Linardi, Julien Longhi

Analysis of Socially Unacceptable Discourse with Zero-shot Learning

Mohamed Rayane GHILENE, Dimitra NIAOURI, Michele LINARDI, Julien LONGHI
ENSEA Engineering School, ETIS UMR-8051 CY Cergy Paris Université, AGORA CY Cergy Paris Université
rayane.ghilene@ensea.fr, {michele.linardi, dimitra.niaouri, julien.longhi} @cyu.fr

Abstract

Socially Unacceptable Discourse (SUD) analysis is crucial for maintaining online positive environments. We investigate the effectiveness of Entailment-based zero-shot text classification (unsupervised method) for SUD detection and characterization by leveraging pre-trained transformer models and prompting techniques. The results demonstrate good generalization capabilities of these models to unseen data and highlight the promising nature of this approach for generating labeled datasets for the analysis and characterization of extremist narratives. The findings of this research contribute to the development of robust tools for studying SUD and promoting responsible communication online. *Accepted for publication in the International Conference on CMC and Social Media Corpora for the Humanities (University Côte d'Azur, France, 2024).*

Keywords: Socially Unacceptable Discourse, Machine Learning, Weakly-Supervised Learning, Explainable Analysis

1. Introduction

Large Language Models (LLMs) have showcased remarkable capabilities in Natural Language Processing (NLP) thanks to their contextual understanding of word embeddings, which have proven to be useful in multiple tasks, including text answering, text generation, and data annotation. LLMs have also shown potential for text classification tasks such as sentiment analysis (Zhang et al., 2023a) by leveraging prompt learning.

In recent years, the spread of Socially Unacceptable Discourse (SUD), including hate speech and toxic comments, in various online platforms has underscored the need for novel tools able to identify and characterize these harmful discourses. However, developing robust automatic SUD classifiers comes with multiple challenges. For instance, the challenge of adopting a universal definition of SUD due to the numerous discourse characterizations causes ambiguity and subjectivity in corpora adopted to train Machine Learning (ML) models (Kocon et al., 2021). Such a scenario poses significant challenges to the creation of well-annotated SUD text corpora that can extensively evaluate the quality of state-of-the-art classification models in large-scale scenarios.

SUD Classification challenges LLMs have obtained state-of-the-art performance in SUD text classification tasks. In this sense, Carneiro et al. (2023) have recently shown that Masked Language Models (MLM) represent a strong candidate classifier option in multiple online annotated corpora. At the same time, Causal Language Models (CLM), which are LLM variants specifically trained to learn cause-effect dynamics (usually adopted by generative AI) can also be successfully leveraged in hate speech classification (Zhang et al., 2023b).

Despite the effectiveness of these models, we note that LLMs lack generalizability in SUD modeling due to their nature, which consists of understanding statistical relationships between words rather than modeling the meaning of these words within their context. Zhang et al. (2023) show that LLMs often obtain solid classification performance in the presence of language stereotypes (e.g., race or religion-related).

On the other hand, in a large-scale context (Carneiro et al., 2023), where heterogeneous subdomains of toxic speech require to be differentiated (i.e., multi-class classification) LLMs are not capable of providing accurate classification due to the presence of overlapping characteristics among different speech classes, but also for the presence of subtle linguistic nuances that require to understand the underlying context to be detected.

Moreover, the annotation schema plays a crucial role in the supervised model training. Often, SUD annotation is subjective and prone to biases resulting from the annotator's background, gender, first language, age, and education (Al Kuwatly et al., 2020). For instance, significant disagreement among annotators from different cultures regarding the offensiveness of online language has been reported in previous studies (Thorn Jakobsen et al., 2022).

Contribution In this work, we present a novel SUD analysis framework, in which we adopt a zero-shot learning paradigm for the automatic detection and characterization of SUD in a large-scale context composed of multiple heterogeneous corpora. Specifically, we leverage natural language inference (NLI) pre-trained models to perform SUD inference (a.k.a. entailment) in text instances. The benefit of this approach is two-fold: first, we do not require data complying with a fixed annotation schema, which may be prone to human bias, second, it will permit to leverage human expertise for hypothesis engineering and validation (Goldzycher and Schneider, 2022), where the users can incorporate their understanding of a specific domain or field to guide the classification process.

2. SUD Framework based on Natural Language Inference

In our solution, we leverage Natural language Inference (NLI) pre-trained models, which are a specific type of NLP models trained to understand the relationship between two pieces of text, namely the *premise* and the *hypothesis* (a new text, potentially related to the premise).

Premise (t)	Hypotheses	Candidate Labels	Entailment Score
what's the difference between a pencil arguing and a woman arguing a pencil has a point	This example is	hate	**0.43**
	This example is	offensive	0.35
	This example is	toxic	0.22

Table 1: Entailment-based zero-shot classification. For every text t (premise) in the dataset, we create multiple hypothesis by considering several known SUD labels.

2.1. Entailment Template

To define premise-hypothesis entailment, we follow a methodology similar to the one proposed for text classification (Gera et al., 2022), adapted to perform unsupervised data labeling.

In this regard, we showcase an illustrative example, drawing inspiration from prior research (Yin et al., 2019) which we have tailored to SUD analysis, as depicted in Figure 1. Here, a hateful premise can be assigned to different labels (hypothesis) according to the perspective under the lens (sentiment, tone of the speech, topics, etc.).

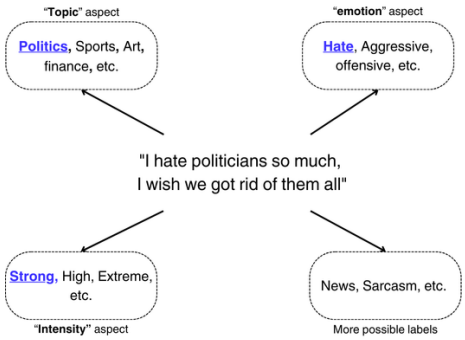

Figure 1: A piece of text can be assigned labels that describe the different aspects of the text. Relevant labels are in blue. Different characterizations of a hateful stance are at the basis of hate speech analysis (Qian et al., 2019).

We thus propose an entailment-based framework, where we couple each premise (text item in a corpus) with a hypothesis stating which class it belongs to. We construct a pair (text item/hypothesis) for each possible SUD class present in the annotation schema of the dataset. Constructed pairs become the input of an NLI model that infers a confidence (entailment) score. In this respect, we consider the output of the softmax layer [1] of an NLI model, where for each hypothesis a probability is assigned between 0 (contradictory hypothesis) and 1.0 (entailed hypothesis).

In Table 1 we report an entailment example that we obtain using a zero-shot learning paradigm to perform an unsupervised premise/hypothesis entailment. Note that a hypothesis is composed of a prefix and a candidate label arbitrarily chosen by the user.

[1] In our case, the softmax layer takes a textual feature vector (learned by the model) of real-valued numbers, transforming it into a probability distribution over a set of possible categories (hypothesis).

2.2. Entailment Models

To perform zero-shot entailment-based text classification on the SUD data, we use models trained specifically for natural language interference (NLI). Such models are pre-trained on the MNLI (Multi-Genre Natural Language Inference) dataset (Williams et al., 2018) which is a large collection of sentence pairs used to evaluate models on their ability to understand entailment between sentences.

It contains over 433,000 sentence pairs in English, drawn from ten different genres of written and spoken text, including news articles, fiction, and conversations. Each pair consists of a premise sentence (source) and a hypothesis sentence (target).

Models trained on the MNLI dataset have the ability to generalize well to different types of textual data, thanks to the diversity of genres they have encountered in the training procedure. For the SUD classification task, we use the following models:

- **Roberta-large-mnli**, BERT (Devlin et al., 2019), which is a transformer-based language model pre-trained on English text using a masked language modeling (MLM) objective and fine-tuned on the Multi-Genre Natural Language Inference (MNLI) corpus.
- **Bart-large-mnli** (Lewis et al., 2020), which is a transformer encoder-decoder (seq2seq) model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder. BART is pre-trained by corrupting text with an arbitrary noising function and learning a model to reconstruct the original text. In our work, we consider the model "facebook/bart-large-mnli", BART version pre-trained on MNLI (Williams et al., 2018) dataset, for Entailment-based Zero shot classification.

  We also consider models trained on other NLI datasets:

- **xlm-roberta-large-xnli-anli**, is a variant of the XLM-RoBERTa architecture proposed in (Conneau et al., 2020), fine-tuned on the XNLI (Cross-lingual Natural Language Inference) (Conneau et al., 2018) and ANLI (Adversarial Natural Language Inference) (Williams et al., 2020) datasets. Its primary application is in cross-lingual natural language inference, which involves determining the relationship (such as entailment, contradiction, or neutrality) between pairs of sentences across multiple languages.
- **MoritzLaurer/mDeBERTa-v3-base-xnli-multilingual-nli-2mil7**, multilingual natural language inference (NLI) model based on the mDeBERTa-v3 architecture, fine-tuned on a combination of the XNLI dataset and an additional multilingual NLI dataset with 2.7 million examples. The mDeBERTa-v3 architecture enhances its performance by incorporating improvements in transformer design, such as disentangled attention and enhanced mask decoder.

3. Empirical Evaluation

To validate our solution, we perform zero-shot entailment-based classification on several publicly available datasets

Dataset	Source	Sample type	# Samples	Labels
Davidson	(Davidson et al., 2017)	Tweets	25,000	hate, offensive, neither
Founta	(Founta et al., 2018)	Tweets	100,000	abusive, hate, neither
Fox	(Gao and Huang, 2017)	Threads	1,528	hate, neither
Gab	(Qian et al., 2019)	Posts	34,000	hate, neither
Grimminger	(Grimminger and Klinger, 2021)	Tweets	3,000	hate, neither
HASOC2019	(Mandl et al., 2019)	Facebook, Twitter posts	12,000	hate, offensive, profane, neither
HASOC2020	(Mandl et al., 2020)	Facebook posts	12,000	hate, offensive, profane, neither
Hateval	(Basile et al., 2019)	Tweets	13,000	hate, neither
Olid	(Zampieri et al., 2019)	Tweets	14,000	offensive, neither
Reddit	(Yuan and Rizoiu, 2022)	Posts	22,000	hate, neither
Stormfront	(De Gibert et al., 2018)	Threads	10,500	hate, neither
Trac	(Kumar et al., 2018)	Facebook posts	15,000	aggressive, neither

Table 2: Summary of datasets (Carneiro et al., 2023)

(Carneiro et al., 2023). Below, we introduce the datasets employed and the results acquired. For the sake of reproducibility, the implemented source code used in the evaluation is publicly available on a public repository [2].

3.1. Datasets

We conducted our evaluation in 12 publicly available datasets containing up to 12 different classes of SUD (Carneiro et al., 2023). In Table 2 we report a detailed overview of the English datasets considered in our study.

3.2. Evaluation of SUD Classifiers

The first goal of our evaluation is to compare the entailment models (unsupervised) with the results we obtain adopting a supervised classifier that has been specifically trained over the annotation schema provided in each dataset.

Such experiment will permit us to answer the question: ***How performance of an elastic and unsupervised method that does not rely on prior SUD knowledge (i.e. the entailment-based zero-shot learning) compare to the ones of a classifier trained over SUD knowledge?*** For this latter, we consider a state-of-the-art MLM, namely BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019)

Note that Masked language models (MLMs), like BERT, are DL models trained to reconstruct masked tokens within the input sequence. Specifically, MLMs learn to predict the original vocabulary identity of a masked word, relying solely on its contextual cues. The significant advantage of those models lies in their bidirectional context, considering both preceding and subsequent tokens during the prediction process. In this work, we fine-tune BERT with the available SUD labels in each evaluated dataset.

We also consider a *shallow* learning baseline such as Logistic Regression (LR) (Grimm and Yarnold, 1995) applied to the numerical representation of tokenized text (text vectorization) [3].

Note that the entailment models we adopt were not pre-trained by leveraging any available ground truth of SUD, and thus they are unsupervised methods in that respect. We base our comparison on the macro F1 score, which is an averaging method for the F1 score that's recommended when working with class imbalance. F1 score is a harmonic mean that combines two performance measures for text classifiers: precision (P) and recall (R). These metrics are computed as follows: $\mathbf{R} = \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$ and $\mathbf{P} = \frac{\text{True Positives}}{\text{True Positives+False Positives}}$.

Hypothesis Testing	roBERTa	BART	mDeBERTa	XLM-roBERTa
this text contains {} speech.	**45.7**	27.6	30.5	**40.9**
this text conveys {} speech.	40.8	34.7	29.6	35.8
this text reflects {} speech.	38.3	35.5	35.3	33.8
this text shows {} speech.	35.1	38.5	27.6	35.7
this text implies {} speech.	33.2	39.6	29.1	32.1
this text reveals {} speech.	37.8	41.6	28.1	32.8
this text exhibits {} speech.	38.8	33.3	24.2	40.4
this text portrays {} speech.	33	36.3	34.6	31.6
this text discusses {} speech.	34.8	37.9	38.9	34.5
this text addresses {} speech.	34.2	38	38.3	37.1
this text illustrates {} speech.	35.9	43	34.2	32.2
this text expresses {} speech.	44.5	35.7	37.3	32.9
this text articulates {} speech.	45.1	42.5	35.8	31
this text suggests {} speech.	30.1	38.6	31.6	32.8
this text narrates {} speech.	43.2	40.5	**38.4**	35.1
this text questions {} speech.	32.6	42	16.4	28.6
this text demonstrates {} speech.	35	42.2	24.7	31.5
this text supports {} speech.	22.6	**44,4**	30.3	31.9
this text has {} speech.	41.1	32.5	12.9	39.3

Table 3: Hypothesis Testing F1 Scores

The F1 score is calculated based on these metrics as $\textbf{F1 Score} = 2 \cdot \frac{\text{P·R}}{\text{P+R}}$. And then the macro F1 score is computed as $\textbf{Macro F1} = \frac{1}{C} \cdot \sum_{i=1}^{C}(F1_i)$, where C is the total number of classes.

3.2.1. Template Selection

In our evaluation, we note that hypothesis construction plays a crucial role in NLI model performance, which has a sensitive and different impact on the considered NLI models, each adopting a different Token masking procedure at the pre-training stage.

In Table 3, we report the hypothesis templates we consider in our work. In detail, we have tested different active parts, i.e., the verb in the formulation, noticing a remarkable impact (+/- 20 in average F1 score) on average SUD classification performance that we report for each model in Table 3. We observe that the four considered NLI models reach the best F1 score using three different hypothesis templates, which we use in the remaining part of the evaluation.

In the same manner, we note that using the word *neither* in

[2] https://github.com/rayaneghilene/ARENAS_Automatic_Extremist_Analysis/tree/main/Entailment_framework

[3] https://keras.io/api/layers/preprocessing_layers/text/text_vectorization/

	Supervised SUD classification		Unsupervised SUD classification (entailment-based)			
Dataset	BERT	LR	Bart-large-mnli	Roberta-large-mnli	xlm-roBERTa	mDeBERTa
Davidson	73	69.5	47.3	44.7	41.5	39.9
Founta	70.1	73.7	57.4	57.5	42.8	36.1
Fox	47.8	69.7	56.1	55.2	52.5	48.7
Gab	87.5	89.0	64.7	67.1	58.3	55.4
Grimminger	51.9	50.4	**52.5**	**56.1**	48.8	38.5
HASOC2019	32.9	39.9	27.5	30.9	17.8	25.8
HASOC2020	41.7	52.5	36.7	**42.7**	20.4	26.4
Hateval	63.6	70.6	59.7	61.4	57.2	54.6
Olid	65.6	71.9	61.6	61.5	52.1	55.5
Reddit	81.7	83.0	56.3	58	50.9	46
Stormfront	66.9	68.4	62	62.6	55.2	51.3
Trac	67.1	69.2	52.1	64.2	61.7	55.5

Table 4: Macro F1 Score (%) of supervised SUD classification VS Entailment-based unsupervised SUD classification with the NLI models.

the hypothesis template does not provide any contextual information to the inference phase of the neutral class, resulting in sensibly low classification performance. We obtain the best performance using the term *neutral speech* in the hypothesis instead of the word *neither* found in each dataset annotation schema (see Table 2).

3.2.2. Results and Discussion

We report experimental results in Table 4. As expected, entailment-based model classification shows slightly lower performance when using entailment models compared to a pre-trained MLM. However, this is not the case for all the datasets, in the Grimminger dataset, our approach outperforms the supervised counterparts, showing a better ability in considering the discourse context at the entailment stage, rather than leveraging correlations among text items in the training set, as in the case of the supervised counterparts.
Furthermore, Roberta-large-mnli and Bart-large-mnli exhibit overall better performance than xlm-roBERTa and mDeBERTa, suggesting that pre-training over the MNLI dataset, which covers a wide range of different spoken and written text is a more suitable choice for SUD analysis.
It is also important to note that such results are similar to the ones obtained by (Gera et al., 2022) when performing zero-shot entailment on other types of text classification. To the best of our knowledge, we are the first to adopt such techniques in SUD analysis.
To conclude, we also observe that there is no clear winner among the supervised classifiers, and a simple Logistic Regression represents an effective solution in the majority of the datasets.

Mitigating biases in the classification To further reduce user bias that may occur in the definition of the hypothesis we adopt GloVe (Pennington et al., 2014) token masking.
This procedure consists of masking tokens highly correlated with the label used in the hypothesis, causing the models to rely on the context provided by the remaining part of the speech in the classification task.
For each text, we mask the tokens with the highest GloVe similarity to the class name following the idea proposed in (Gera et al., 2022).
For example, when classifying *offensive* SUD, the words correlated to offensive language will be masked in the text.
The experimental results reported in Table 5 show that the effect of token masking comes only with a slight performance decrease (in most datasets) compared to the results obtained by the entailment models off-the-shelf. Such results suggest how entailment-based SUD classification can not only leverage class stereotypes, but it can potentially leverage the remaining part of the speech.

Dataset	Bart-large-mnli	Bart-large-mnli + Mask	RoBERTa-large-mnli	RoBERTa-large-mnli + Mask
Davidson	47.3	40.3	44.7	42.5
Founta	57.4	53	57.5	49.8
Fox	56.1	55.5	55.2	57
Gab	64.7	61.4	67.1	66.6
Grimminger	52.5	50.5	56.1	56.4
HASOC2019	27.5	23.3	30.9	29.8
HASOC2020	36.7	28.6	42.7	37.3
Hateval	60.8	58.6	61.4	61.3
Olid	61.6	59.5	61.5	61.8
Reddit	56.3	53.6	58	59.8
Stormfront	62	59.1	62.6	62.6
Trac	52.1	47.6	64.2	63.4

Table 5: **Zero-shot text classification with token masking** For each zero-shot entailment model and dataset, we compare the macro F1 score of the off-the-shelf model to its score when performing token masking.

4. Conclusion and Future Work

This paper investigates the effectiveness of zero-shot entailment using NLI models for SUD classification.
Through preliminary experimentation, these models showcased generalization capabilities comparable with supervised counterparts. Such a scenario highlights the entailment-based model's potentiality to exploit contextual information in the text rather than learning intra-class correlation using a fixed annotation schema, which may be sensitive to stereotypes of certain kinds of SUD.
The preliminary results we obtained motivate several future work directions. First, we would like to explore how to effectively learn templates that allow linguists to use semantically richer and unstructured annotation schemes, also studying scalability issues and tradeoffs of large entailment hypothesis spaces. We believe that such capability can support supervised learning models currently adopted

in SUD analysis to reduce the impact of annotator bias and sensitivity to class stereotypes.

This result will be a valuable advance for the CMC corpora community and work in corpus linguistics, allowing synergies between AI and corpus linguistics researchers.

5. References

Al Kuwatly, H., Wich, M., and Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*.

Basile, V., Bosco, C., Fersini, E., et al. (2019). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*.

Carneiro, B. M., Linardi, M., and Longhi, J. (2023). Studying socially unacceptable discourse classification (SUD) through different eyes: "are we on the same page ?". *CoRR*, abs/2308.04180.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., and Stoyanov, V. (2018). Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., GuzmÃ¡n, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale.

Davidson, T., Warmsley, D., Macy, M. W., et al. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM*. AAAI Press.

De Gibert, O., Perez, N., García-Pablos, A., et al. (2018). Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Devlin, J., Chang, M., Lee, K., et al. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: NAACL-HLT*.

Founta, A., Djouvas, C., Chatzakou, D., et al. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM*.

Gao, L. and Huang, R. (2017). Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP*.

Gera, A., Halfon, A., Shnarch, E., et al. (2022). Zero-shot text classification with self-training. In *Proceedings of EMNLP*.

Goldzycher, J. and Schneider, G. (2022). Hypothesis engineering for zero-shot hate speech detection. *arXiv preprint arXiv:2210.00910*.

Grimm, L. G. and Yarnold, P. R. (1995). *Reading and understanding multivariate statistics.* American psychological association.

Grimminger, L. and Klinger, R. (2021). Hate towards the political opponent: A twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EACL 2021*.

Kocon, J., Figas, A., Gruza, M., et al. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Inf. Process. Manag.*, 58(5).

Kumar, R., Reganti, A. N., Bhatia, A., et al. (2018). Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402*.

Lewis, M., Liu, Y., Goyal, N., et al. (2020). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*.

Mandl, T., Modha, S., Majumder, P., et al. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*.

Mandl, T., Modha, S., Kumar M, A., et al. (2020). Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of the 12th annual meeting of the forum for information retrieval evaluation*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

Qian, J., Bethke, A., Liu, Y., et al. (2019). A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of EMNLP-IJCNLP*.

Thorn Jakobsen, T. S., Barrett, M., Søgaard, A., et al. (2022). The sensitivity of annotator bias to task definitions in argument mining. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*.

Williams, A., Nangia, N., and Bowman, S. R. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: NAACL-HLT*.

Williams, A., Thrush, T., and Kiela, D. (2020). Anlizing the adversarial natural language inference dataset. *arXiv preprint arXiv:2010.12729*.

Yin, W., Hay, J., and Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of EMNLP-IJCNLP*.

Yuan, L. and Rizoiu, M. (2022). Detect hate speech in unseen domains using multi-task learning: A case study of political public figures. *CoRR*, abs/2208.10598.

Zampieri, M., Malmasi, S., Nakov, P., et al. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Zhang, P., Chai, T., and Xu, Y. (2023a). Adaptive prompt learning-based few-shot sentiment analysis. *Neural Process. Lett.*, 55(6).

Zhang, Z., Chen, J., and Yang, D. (2023b). Mitigating biases in hate speech detection from A causal perspective. In *Findings of the Association for Computational Linguistics: EMNLP*.