Comparative Study of Different Transformer Models for Hope Speech Detection

Luis Ramos[1,*,†], Hiram Calvo[1] and Olga Kolesnikova[1]

[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, Mexico City, Mexico

Abstract

Through artificial intelligence and natural language processing, online phenomenons like hate speech, cyberbullying, and fake news can be analysed on social media platforms due to the extensive user engagement they capture and the extensive use of transformer architecture. Detection of hope speech, which goes beyond positivism and self-motivating in a goal-directed manner, is a new and more impactful approach that fosters user well-being. The solution proposed in this paper is based on the idea of using pre-trained transformers in tasks of detecting language associated with intentions, such as hate speech, toxic language, depressive speech, or speech expressing feelings; Using fine-tuning technique is possible to adapt this patterns from negative to positive speech detection, due to the diverse type of speech added for this dataset, like sarcasm as negative speech. As for our evaluation on different pretrained transformers, a RoBERTa model performed well in this task on the development set with a macro-F1 of 0.830 and accuracy of 0.818, while other models like XML-based recorded the highest recall of 0.933, but a RoBERTa model also recorded the highest precision of 0.816. On the test set, the RoBERTa-based models performed best, achieving a Macro F1 score of 0.84. This indicates a strong performance on this task, even with no prior training on hope detection datasets.

Keywords

Hope Speech, Transformers, Fine-Tuning, NLP, Classification

1. Introduction

Social media platforms serve as a growing repository of collective information with the activities and interactions of millions of users, and these benefits allow social phenomena to be discovered and analyzed [1]. These platforms exhibit social phenomena, for example hate speech [2], fake news [3], threat [4] and cyberbullying [5], which have been studied in the context of Artificial Intelligence (AI) and Natural Language Processing (NLP). Recently, transformer models have rapidly become the dominant architecture for natural language processing [6, 7].

Hope is defined as the ability to build methods that are likely to assist in achieving goals, and self-motivate through agentic reasoning to put such methods into action [8, 9]. Hope should not be mistaken with a good attitude or mere positivity because it is much more than that as a feeling. Examining the notion of hope sharpens the understanding of the goals and anticipations of people, cultures, and even their subdivisions, which may include sexes and races[10]. Hope speech detection is a recent task in natural language processing (NLP) [11, 12, 13, 14]. This task is fuelled by opportunities in social media, where the detection of hopeful expressions can enhance user well-being [12].

In this paper, we tested diverse transformer models [15, 16] for binary hope-speech detection in English language at IberLEF 2025 [14, 17, 18], resulting in a comparison of classifiers performance. The idea behind of using pre-trained transformers in tasks of detecting language associated with intentions (such as hate speech, toxic language, depressive speech, or speech expressing feelings) is the use of fine-tuning technique to adapt this patterns (positive/negative speech) to hope speech detection.

The leading model was *cardiffnlp/twitter-roberta-base-hate-latest*, where the macro-F1 score was at 0.830 and accuracy at 0.818. The model *textdetox/xlmr-large-toxicity-classifier-v2* had the high-

IberLEF 2025, September 2025, Zaragoza, Spain

[*] Corresponding author.

[†] These authors contributed equally.

lramos2020@cic.ipn.mx (L. Ramos); hcalvo@cic.ipn.mx (H. Calvo); kolesnikova@cic.ipn.mx (O. Kolesnikova)

0009-0008-5586-6668 (L. Ramos); 0000-0003-2836-2102 (H. Calvo); 000-0002-1307-1647 (O. Kolesnikova)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2025-12-10

Text	Label
#USER# Anyway love u bubbly i know i can count on you when its about fairy tail	Hope
Crawling into #USER# bed for a quick nap...here's hoping the lass doesn't mind too much..	Hope
Find me a mass shooting in Utah where concealed carry on campus is legal.	Not Hope
#USER# You're not a bad person at all. Just can't believe you would second guess yourself.	Not Hope

Table 1
Data Samples

est recall of 0.933 but had lower precision and accuracy compared to other models. The model *rafalposwiata/deproberta-large-depression* dominated in precision scoring at 0.816 and also achieving a solid F1 score of 0.814. Other models such as *papluca/xlm-roberta-base-language-detection*, *textdetox/bert-multilingual-toxicity-classifier* and *clapAI/roberta-large-multilingual-sentiment* had balanced results at roughly 0.79-0.81 F1 score and around 0.79 in accuracy, with other models trailing slightly behind these benchmarks.

These results reaffirm that transformers models performing well, which capture more deeply the complex semantic structures of hope expressions, enabling enhanced precision and recall in hope-speech recognition.

2. Literature Review

Hope speech detection is a novel area of research, and in this section we provide a review of different approaches in binary as well as multiclass classification datasets and techniques [19, 20]. This review aims to summarize the progress made so far and highlight the existing gaps in hope speech detection.

Recent years have seen an increasing interest in the computational analysis of hope speech. Bharathi Chakravarthi,[21] provided the first multilingual hope-speech corpus within the Equality, Diversity, and Inclusion framework, which includes 28 451 English, 20 198 Tamil, and 10 705 Malayalam YouTube comments. Comments were marked up with positive content encouraging annotations representing different languages and cultures. The author experimented with various traditional machine learning models; however, all of them exhibited low performance.

Daniel García-Baena et al.,[22] created a dataset of 1,650 tweets divided into two categories: Hope Speech and Non-Hope Speech. Through their analysis, they showcased that a fine-tuned BETO surpassed both traditional (SVM, Naive Bayes, and Logistic Regression) and neural-based classifiers (MLP, CNN, and BiLSTM) when utilizing BERT-level embeddings with an F1-score of 85.12%.

Fazlourrahman Balouchzahi et al.,[12] recently proposed PolyHope, an English tweet corpus with a two-level classification of Hope, binary and a finer three-class subdivision into Generalized, Realistic, and Unrealistic Hope. Authors reported that transformers achieved a macro F1 score of 0.85 on the binary and 0.72 for the multiclass classification. In their subsequent work on Urdu language[23], created a five-class dataset that highlighted model divergence: Logistic Regression achieved the best for binary classification with a macro F1 score of 0.7593 and for multiclass classification RoBERTa was the best with a macro F1 score of 0.4801.

Grigori Sidorov et al.,[13] brings forth two key contributions: building the first multiclass hope-speech corpus for Spanish and German based on tweets, and accomplishing a thorough assessment of detection frameworks across traditional machine learning, deep learning, and transformer models. Fine-tuned XLM-RoBERTa-base achieved an F1 score of 0.6801 on the Spanish data while UKLFR/GottBERT-base achieved 0.6977 on the German data, widely surpassing all non-transformer models.

Recently, shared tasks have contributed to the evaluation methods for sharpening models in multilingual natural language processing. The detection of hope speech within the LT-EDI 2022 shared task [24] sought to classify "hope speech" versus "non-hope speech" in five typologically divergent languages: Tamil, Malayalam, Kannada, English, and Spanish, thus advancing the development of models with translingual and transcultural inclusivity. Likewise at IberLEF 2023 shared task [25], HOPE: Multilingual Hope Speech Detection, consisted of identifying whether texts written in English or Spanish contained

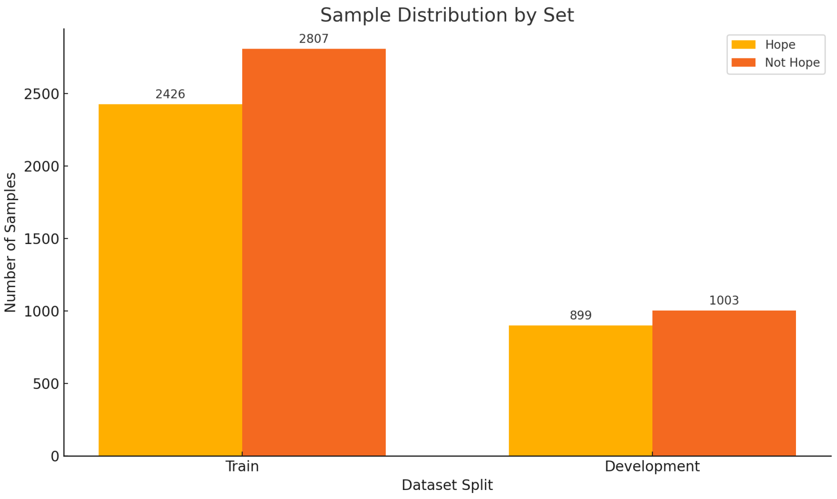

Figure 1: Data Statistics

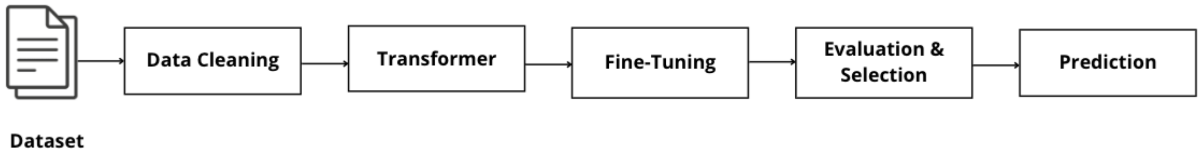

Figure 2: Overview of Methodology

hope speech. In 2024 was released the second edition of this shared task on multilingual hope speech detection, HOPE 2024 [26], which was held within the framework of the IberLEF 2024.

Although all of these proposals contribute towards a coherent and important trend in the computational analysis of hope speech, they also suggest contexts wherein this analysis can be deepened and further developed.

3. Data Description

Dataset was collected and split by the organizers of PolyHope shared task [14]. The dataset contains English tweets and its statistics are presented in Figure 1, noting there exist an imbalance between classes. Computing the imbalance ratio (IR) as in Equation 1, the IR of this dataset is 2.75. An increased imbalance ratio IR results in a greater extent imbalance of the dataset, making it harder to classify datasets with higher IR [27]. Table 1 shows some samples from the dataset. The predictions were generated on the test set, which contains 2065 texts.

$$IR = \frac{N_{maj}}{N_{min}} \tag{1}$$

4. Methodology

This section outlines the specifics of the approach taken with different transformer models. Only one data cleaning phase is needed for this method before the tweets are input into transformer models. Cleaning phase and transformer model details are described in detail in the following subsections, as each phase of the proposed methodology showed in Figure 2.

Model	Macro Precision	Macro Recall	Macro F1	Acc
papluca/xlm-roberta-base-language-detection	0.807	0.775	0.791	0.783
facebook/roberta-hate-speech-dynabench-r4-target	0.760	0.881	0.816	0.790
textdetox/xlmr-large-toxicity-classifier-v2	0.665	0.933	0.777	0.717
cardiffnlp/twitter-roberta-base-hate-latest	0.814	0.847	0.830	0.818
rafalposwiata/deproberta-large-depression	0.816	0.812	0.814	0.804
fatmhd1995/toxic_comment_model_ethos_ft	0.761	0.835	0.796	0.775
textdetox/bert-multilingual-toxicity-classifier	0.788	0.840	0.813	0.797
DunnBC22/bert-large-uncased-Hate_Offensive_or_Normal_Speech	0.761	0.838	0.798	0.776
clapAI/roberta-large-multilingual-sentiment	0.788	0.841	0.814	0.797

Table 2
Model performances on development set

Model	Weighted Precision	Weighted Recall	Weighted F1	Macro Precision	Macro Recall	Macro F1	Acc
textdetox/xlmr-large-toxicity-classifier-v2	0.82	0.81	0.81	0.82	0.80	0.81	0.81
rafalposwiata/deproberta-large-depression	0.84	0.84	0.84	0.84	0.84	0.84	0.84
cardiffnlp/twitter-roberta-base-hate-latest	0.84	0.84	0.84	0.84	0.84	0.84	0.84
facebook/roberta-hate-speech-dynabench-r4-target	0.84	0.84	0.84	0.84	0.84	0.84	0.84

Table 3
Model performances on test set

4.1. Data Cleaning

This phase encompasses lowercase and the removal of emojis, URLs, numerals, special symbols, words framed by parentheses or number symbols (as exemplified in Table 1), as well as stop words (using NLTK library). Additionally, lemmatization of words was applied utilizing Spacy library. This phase prepares the text for further analysis and modelling [28].

4.2. Transformer Models

Transformer architectures utilize self-attention mechanisms to dynamically capture context at different levels [29]. In hope speech detection, they have achieved promising results identifying hope expressions [30]. Diverse transformers models from Hugging Face[1] were selected for this proposal, *papluca/xlm-roberta-base-language-detection*, *textdetox/xlmr-large-toxicity-classifier-v2*, *cardiffnlp/twitter-roberta-base-hate-latest*, *rafalposwiata/deproberta-large-depression*, *fatmhd1995/toxic_comment_model_ethos_ft*, *textdetox/bert-multilingual-toxicity-classifier*, *DunnBC22/bert-large-uncased-Hate_Offensive_or_Normal_Speech*, *clapAI/roberta-large-multilingual-sentiment*, and *facebook/roberta-hate-speech-dynabench-r4-target*, which was trained for hate speech detection [31]. Every model was trained using the following specific parameter values and the rest in default: 5 epochs, train and development (eval) batch size of 16, *load_best_model_at_end = True* and *metric_for_best_model='f1'*. The last two parameters selects automatically the best model based on the best F1 score.

5. Results

For the development set, we offer only macro metrics, since the primary comparative performance metric is based on macro F1 score. However, for the best model's performance on the test set, we also report macro precision, macro recall, weighted precision, weighted recall, weighted F1 score, and accuracy. Combined, these metrics provide a holistic assessment of the model's performance.

[1] https://huggingface.co/

The Table 2 shows the comparison of the models performance on development set. During development, the *cardiffnlp/twitter-roberta-base-hate-latest* model reached a Macro F1 of 0.830, accompanied by scores of 0.814 for precision and 0.847 for recall; those figures signal a solid, if not perfect, equilibrium between the two main classes. The *textdetox/xlmr-large-toxicity-classifier-v2*, by contrast, snared a Macro recall as high as 0.933, flagging almost every instance of hope speech, yet it did so at the cost of lower precision, which bottomed out at 0.665. The *rafalposwiata/deproberta-large-depression* variant drew attention for its relatively high precision of 0.816, trimming the number of false positives and yielding a Macro F1 of 0.814 alongside a recall of 0.812.

The Table 3 shows the comparison of the models' performance on test set. When the models were evaluated on the hold-out test set, the *textdetox/xlmr-large-toxicity-classifier-v2*, the *rafalposwiata/deproberta-large-depression* architecture, and the *cardiffnlp/twitter-roberta-base-hate-latest* implementation wound up reporting nearly identical weighted and macro statistics: F1 scores landed in a narrow band from 0.81 to 0.84, and accuracy brushed the same range. That clustering of results hints that the disparate pre-training regimens of these RoBERTa- and DeBERTa-derived systems translate into comparable real-world robustness once the data are unseen.

6. Discussion

Pre-trained networks that have already been exposed to linguistically similar genres yield surprisingly robust semantic and pragmatic footprints, allowing a system to spot hope speech without the crippling overhead of retraining from scratch. The practical dilemma then narrows to a simple trade-off: *textdetox/xlmr-large-toxicity-classifier-v2* favours volume-so the developer who prioritizes sheer message count works almost by instinct-while *rafalposwiata/deproberta-large-depression* tightens the aperture and lightly hoists precision at the expense of recall. That imbalance in the source data urges testers to center the evaluation on macro-averaged metrics and to experiment with resampling or cost-sensitive routines if the generalization is to survive outside the original dataset. Even so, the figures live almost entirely in English Twitter, meaning any real deployment will still need cross-lingual trials and perhaps a few heuristics for picking up quieter pragmatic cues.

7. Conclusion

Recent experiments show that the studied Transformer architectures can flag hopeful language with fidelity, even though none were purpose-built for this niche. Macro F1 measures surpassing 0.81 on both the dev and held-out sets underscore the strength of their embedded contexts. A noticeable tension sits between precision and recall: the *textdetox/xlmr-large-toxicity-classifier-v2* pipeline opts to capture nearly every instance of hope speech, while *rafalposwiata/deproberta-large-depression* trades off by issuing fewer, but more certain, affirmative calls. Still, a persistent class imbalance-ratio index clocking in at 2.75-and a test set barely nudging past 2,000 samples place a ceiling on how widely the numbers can be extrapolated. Broadening the corpus could either bolster these already solid scores or expose latent fragilities in model behaviour.

Declaration on Generative AI

During the preparation of this work, the authors used GPT in order to: Grammar and spelling check.

References

[1] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, P. Spyridonos, Community detection in social media: Performance and application considerations, Data mining and knowledge discovery 24 (2012) 515–554.

[2] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020).

[3] P. M. Subhash, D. Gupta, S. Palaniswamy, M. Venugopalan, Fake news detection using deep learning and transformer-based model, in: 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2023, pp. 1–6.

[4] S. Butt, M. Amjad, F. Balouchzahi, N. Ashraf, R. Sharma, G. Sidorov, A. Gelbukh, Emothreat@ fire2022: Shared track on emotions and threat detection in urdu, in: Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation, 2022, pp. 1–3.

[5] W. Tapaopong, A. Charoenphon, J. Raksasri, T. Samanchuen, Enhancing cyberbullying detection on social media using transformer models, in: 2024 5th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON), IEEE, 2024, pp. 1–5.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[7] O. Kolesnikova, M. Shahiki Tash, Z. Ahani, A. Agrawal, R. Monroy, G. Sidorov, Advanced machine learning techniques for social support detection on social media, Heliyon 11 (2025) e43437. URL: https://www.sciencedirect.com/science/article/pii/S2405844025018225. doi:https://doi.org/10.1016/j.heliyon.2025.e43437.

[8] C. R. Snyder, Hope theory: Rainbows in the mind, Psychological inquiry 13 (2002) 249–275.

[9] M. Shahiki-Tash, J. Armenta-Segura, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma at hope2023iberlef: Hope speech detection using lexical features and convolutional neural networks, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org, 2023.

[10] F. Balouchzahi, S. Butt, G. Sidorov, A. Gelbukh, Cic@ lt-edi-acl2022: Are transformers the only hope? hope speech detection for spanish and english comments, in: Proceedings of the second workshop on language technology for equality, diversity and inclusion, 2022, pp. 206–211.

[11] G. Sidorov, F. Balouchzahi, S. Butt, A. Gelbukh, Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets, Applied Sciences 13 (2023) 3983.

[12] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078.

[13] G. Sidorov, F. Balouchzahi, L. Ramos, H. Gómez-Adorno, A. Gelbukh, Mind-hope: Multilingual identification of nuanced dimensions of hope (2024).

[14] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org, 2025.

[15] M. S. Tash, L. Ramos, Z. Ahani, R. Monroy, H. Calvo, G. Sidorov, et al., Online social support detection in spanish social media texts, arXiv preprint arXiv:2502.09640 (2025).

[16] M. S. Tash, Z. Ahani, O. Kolesnikova, G. Sidorov, Analyzing emotional trends from x platform using senticnet: A comparative analysis with cryptocurrency price, arXiv preprint arXiv:2405.03084 (2024).

[17] S. Butt, F. Balouchzahi, M. Amjad, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Overview of polyhope at iberlef 2025: Optimism, expectation or sarcasm?, Procesamiento del Lenguaje Natural (2025).

[18] S. Butt, F. Balouchzahi, A. I. Amjad, M. Amjad, H. G. Ceballos, S. M. Jimenez-Zafra, Optimism, expectation, or sarcasm? multi-class hope speech detection in spanish and english, arXiv preprint arXiv:2504.17974 (2025).

[19] M. Arif, M. Shahiki Tash, A. Jamshidi, F. Ullah, I. Ameer, J. Kalita, A. Gelbukh, F. Balouchzahi, Analyzing hope speech from psycholinguistic and emotional perspectives, Scientific reports 14 (2024) 23548.

[20] Z. Ahani, M. S. Tash, M. Tash, A. Gelbukh, I. Gelbukh, Multiclass hope speech detection through transformer methods, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS. org, 2024.

[21] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: M. Nissim, V. Patti, B. Plank, E. Durmus (Eds.), Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.

[22] D. García-Baena, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish: The lgbt case, Language Resources and Evaluation 57 (2023) 1487–1514.

[23] F. Balouchzahi, S. Butt, M. Amjad, G. Sidorov, A. Gelbukh, Urduhope: Analysis of hope and hopelessness in urdu texts, Knowledge-Based Systems 308 (2025) 112746.

[24] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, et al., Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the second workshop on language technology for equality, diversity and inclusion, 2022, pp. 378–388.

[25] S. M. Jiménez-Zafra, M. Á. Garcia-Cumbreras, D. García-Baena, J. A. Garcia-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of hope at iberlef 2023: Multilingual hope speech detection, Procesamiento del lenguaje natural 71 (2023) 371–381.

[26] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. García-Cumbreras, A. L. Tonja, J. A. García-Díaz, S. Bozkurt, B. R. Chakravarthi, H. G. Ceballos, R. Valencia-García, et al., Overview of hope at iberlef 2024: Approaching hope speech detection in social media from two perspectives, for equality, diversity and inclusion and as expectations, Procesamiento del lenguaje natural 73 (2024) 407–419.

[27] R. Zhu, Y. Guo, J.-H. Xue, Adjusting the imbalance ratio by the dimensionality of imbalanced data, Pattern Recognition Letters 133 (2020) 217–223.

[28] S. Z. Ridoy, J. Sultana, Z. F. Ria, M. A. Uddin, M. H. Rahman, R. M. Rahman, An efficient text cleaning pipeline for clinical text for transformer encoder models, in: 2024 IEEE 12th International Conference on Intelligent Systems (IS), IEEE, 2024, pp. 1–9.

[29] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, G. Sidorov, Detection of depression severity in social media text using transformer-based models, Information 16 (2025) 114.

[30] M. Krasitskii, O. Kolesnikova, L. C. Hernandez, G. Sidorov, A. Gelbukh, Unveiling hope in social media: A multilingual approach using bert (2024).

[31] B. Vidgen, T. Thrush, Z. Waseem, D. Kiela, Learning from the worst: Dynamically generated datasets to improve online hate detection, arXiv preprint arXiv:2012.15761 (2020).