LeaningTower@LT-EDI-ACL2022: When Hope and Hate Collide

Arianna Muti[♠,∗], **Marta Marchiori Manerba**[♣,∗],
Katerina Korre[♠] and **Alberto Barrón-Cedeño**[♠]
[♠] DIT, Alma Mater Studiorum–Universitá di Bologna, Forlì, Italy
[♣] Universitá di Pisa, Pisa, Italy
{arianna.muti2,aikaterini.korre2,a.barron}@unibo.it
marta.marchiori@phd.unipi.it
∗ A. Muti and M. Marchiori Manerba contributed equally to this work. They should both be considered first authors.

Abstract

The 2022 edition of LT-EDI proposed two tasks in various languages. Task$_{\text{hope}}$ required models for the automatic identification of hopeful comments for equality, diversity, and inclusion. Task$_{\text{antiLGBT}}$ focused on the identification of homophobic and transphobic comments. We targeted both tasks in English by using reinforced BERT-based approaches. Our core strategy aimed at exploiting the data available for each given task to augment the amount of supervised instances in the other. On the basis of an active learning process, we trained a model on the dataset for Task i and applied it to the dataset for Task j to iteratively integrate new silver data for Task i. Our official submissions to the shared task obtained a macro-averaged F_1 score of 0.53 for Task$_{\text{hope}}$ and 0.46 for Task$_{\text{antiLGBT}}$, placing our team in the third and fourth positions out of 11 and 12 participating teams respectively.

1 Introduction

In recent years, many episodes of violence against homosexuals and transsexuals have been observed online (e.g., in YouTube comments[1]) and offline, which escalated into the death of 375 transgender people in 2021 alone.[2] Most of the victims were Black and Latin women, especially sex workers, a fact that highlights the intersection between misogyny, racism, xenophobia and hate towards sex workers. That is why identifying such behaviours online is timely, as it can contribute to limiting the spread of hate. In this regard, two different tasks have been proposed in LT-EDI in various languages:

Homo/Transphobia Detection (Task$_{\text{antiLGBT}}$) Classify a YouTube comment into homophobic, transphobic or non-anti-LGBT content (Chakravarthi et al., 2022b).

Hope Speech Detection (Task$_{\text{hope}}$) Classify a YouTube comment into hope speech or non-hope speech (Chakravarthi et al., 2022a).

We approach both tasks, addressing the English language only.[3] We experiment with two different approaches for Task$_{\text{hope}}$ and four for Task$_{\text{antiLGBT}}$. We aim at augmenting the data to cope with the heavy imbalance in the datasets. All models are built on top of BERT (Devlin et al., 2019). For Task$_{\text{hope}}$ we implement a binary classifier which is our baseline, and we augment data through an active learning approach (Hino, 2020). For Task$_{\text{antiLGBT}}$ we implement a multi-class classifier as our baseline. Then, we augment training data according to three approaches:

- augmenting transphobic instances by adding Tamil data translated into English;
- augmenting non-anti-LGBT content instances by integrating hope speech instances from Task$_{\text{hope}}$; and
- Performing an active learning approach.

The rest of the paper is structured as follows. Section 2 provides an overview of definitions and related work in the field of abusive language detection, focusing in particular on homophobia, transphobia (and hope speech). Section 3 explores the two datasets provided by the shared task. Section 4 describes our models for both tasks and Section 5 outlines the hyperparameters and preliminary experiments. Section 6 presents and discusses our results. Finally, Section 7 draws conclusions.

[1] https://www.bbc.com/news/technology-50166900

[2] https://www.forbes.com/sites/jamiewareham/2021/11/11/375-transgender-people-murdered-in-2021-deadliest-year-since-records-began/

[3] Our implementation is available at https://github.com/TinfFoil/leaningtower_ltedi22.

Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 306 - 311
May 27, 2022

2 Background

The importance of the automatic detection of abusive language has increased together with the popularity of social media (Fortuna and Nunes, 2018). The online discourse often has hateful and offensive connotations towards minorities. The exposure to hate speech can trigger polarization, isolation, depression, and other psychological trauma (Kiritchenko et al., 2021). Becoming aware of this serious societal issue, online platforms have assumed the responsibility of examining and removing hateful posts (Fortuna and Nunes, 2018). Due to the continuous flow of large amounts of contents through social media, hatred is flagged through automatic methods along with human monitoring (Poletto et al., 2021).

In order to foster the development of automatic models for the identification of different kinds of hate speech, diverse supervised datasets and models have been developed. Chakravarthi et al. (2021) proposed a dataset with homophobic and transphobic contents from YouTube, gathering comments from famous YouTubers that raise awareness on the LGBT+ community and also from channels that report pranks and jokes about homosexuals and transsexuals. Given the sensitivity of the topics covered in the videos, the comments posted can often have abusive, offensive or denigratory connotations towards the LGBT+ community. They found out that a combination of machine learning models, including random forests (Breiman, 2001) reinforced with BERT embeddings (Devlin et al., 2019), obtains the best result.

Hope speech, on the other hand, lies on the other end of the spectrum of digital rhetoric. In contrast to hateful comments, a hopeful discourse is characterized by a friendly tone and an intention to inspire, support, include, and encourage members of minorities, who are often subject to judgment, isolation, and suffering (Chakravarthi, 2020). Focusing on spotting hopeful rather than hateful contents offers a twist that seeks to produce a better online ecosystem by promoting rather than limiting comments and opinions.

This angle was explored within the hope speech detection shared task (Chakravarthi, 2020) on HopeEDI, a multilingual collection of YouTube comments.[4] According to Chakravarthi and Muralidaran (2021), the best approach for English achieved 0.93 F_1score: the winning team fine-tuned RoBERTa (Liu et al., 2019) on the three datasets, i.e., the collections in English, Tamil, and Malayalam.

[4] https://sites.google.com/view/lt-edi-2021/home

	train	test
homophobic	215	61
transphobic	8	5
non-anti-LGBT	3,730	924

Table 1: Statistics of the English corpus for $\text{Task}_{\text{antiLGBT}}$.

	train	test
hope speech	2,234	-
non-hope speech	23,347	-

Table 2: Statistics of the English corpus for $\text{Task}_{\text{hope}}$.

Relevant work in this area includes also the contribution of Palakodety et al. (2020), where the authors collect another hope speech dataset of YouTube comments posted on videos related to the India–Pakistan conflict and apply active learning as well to tackle the imbalanced distribution.

3 Datasets

Here, we briefly describe the datasets for $\text{Task}_{\text{antiLGBT}}$ and $\text{Task}_{\text{hope}}$.

$\text{Task}_{\text{antiLGBT}}$ The collection consists of comments of YouTube videos that were annotated by LGBT+ community members. Table 1 shows statistics. The distribution is heavily skewed, with less than 10% of homophobic instances and only 8 instances of transphobia. This low amount of instances could significantly impact a model's capability of spotting transphobic comments.

$\text{Task}_{\text{hope}}$ Table 2 shows statistics for the $\text{Task}_{\text{hope}}$ dataset.[5] Once again, the corpus is heavily imbalanced: only 10% of the instances belong to the hopeful class. As claimed by Chakravarthi (2020), this class distribution reflects a real-world scenario.

4 Systems Overview

In the following paragraphs, we first describe the active learning approach. We then present the specific strategies developed for $\text{Task}_{\text{antiLGBT}}$ and $\text{Task}_{\text{hope}}$ respectively. For $\text{Task}_{\text{antiLGBT}}$, we

[5] The numbers for the test set will be included upon release of the gold labels.

trained four alternative models to identify the best possible configuration: baseline, baseline augmented with Tamil data translated to English, baseline augmented with hope speech data remapped as non-anti-LGBT content and baseline with augmented data from $\text{Task}_{\text{hope}}$ through an active learning approach. For $\text{Task}_{\text{hope}}$, we trained two alternative models: the baseline and the active learning approach.

Cross-task data augmentation through active learning The two tasks at hand are related, as the labels of both datasets can be traced back to hateful and non-hateful instances. Instances of homo/transphobic and hope speech messages can be remapped to their non-hope speech and non-anti-LGBT comments respectively. On the contrary, it is not always true that a non-hope speech instance is homo/transphobic and that a non-anti-LGBT content contains hope speech. Therefore, given the small amount of training instances available for both $\text{Task}_{\text{antiLGBT}}$ and $\text{Task}_{\text{hope}}$, we aim to take advantage of both datasets proposing an approach to augment the training sets for each task. We first add the homo/transphobic and hope speech instances in bulk, and then we filter the uncertain ones, i.e., non-hope speech for $\text{Task}_{\text{antiLGBT}}$ and non-anti-LGBT content for $\text{Task}_{\text{hope}}$, through an active learning approach (Hino, 2020) as follows. Let D_i and D_j be the supervised datasets for both tasks. (*i*) Train model m_i on D_i. (*ii*) Predict the instances in D_j with m_i. (*iii*) Rank the instances in D_j according to the confidence of the prediction score returned by m_i. (*iv*) Transfer the top-k instances in D_j as silver data to D_i. This process is repeated until $|D_j| = \emptyset$ and the final model for Task i is then used to predict on the dev set for Task i.

Specifically, we augment the dataset for $\text{Task}_{\text{hope}}$ by adding in bulk homophobic and transphobic instances remapped to non-hope speech instances. We do the same for $\text{Task}_{\text{antiLGBT}}$ by adding in bulk hope speech instances to non-anti-LGBT content. Then, we use an active learning approach to identify which non-anti-LGBT instances contain hope speech, and which non-hope speech instances contain homophobia/transphobia. In the end we integrate the identified instances (i.e., hope speech and homo/transphobic) in both datasets. Figure 1 represents the approach for $\text{Task}_{\text{hope}}$. First, homophobic and transphobic instances from $\text{Task}_{\text{antiLGBT}}$ are added as non-hope speech. Then, we feed non-anti-LGBT instances

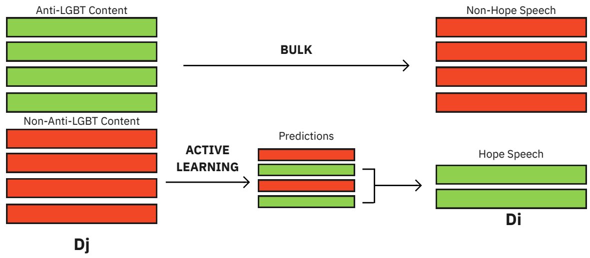

Figure 1: Our strategy to augment the dataset for $\text{Task}_{\text{hope}}$. Anti-LGBT content includes both homophobic and transphobic instances.

to the model trained on $\text{Task}_{\text{hope}}$ dataset. Those which are predicted as hope speech are integrated in the training set. We adopt the same approach for $\text{Task}_{\text{antiLGBT}}$.

4.1 $\textbf{Task}_{\textbf{antiLGBT}}$

Baseline In our first and simplest approach we adopt a similar architecture for both tasks. The model is built on top of BERT (Devlin et al., 2019) with a softmax activation function in the output. For $\text{Task}_{\text{antiLGBT}}$, we adopt a multi-class approach with mutually exclusive categories with three output units. This approach is based on the top-performing model (Muti and Barrón-Cedeño, 2020) at the AMI shared task on the identification of misogynous and aggressive tweets (Elisabetta Fersini, 2020). No external data is considered in this model.

Baseline augmented with Tamil data Whereas we focus on the English language for both tasks, we exploit the provided dataset in Tamil by translating it into English using the GoogleTrans API.[6] One of the main purposes of this cross-language augmentation was increasing through machine translation the amount of transphobic instances with the 155 available in Tamil. However, only some of them were successfully translated, as many of the sentences remained in Tamil, therefore we could only exploit 54 instances.

Baseline augmented with hope speech data A first cross-task data augmentation involved adding in bulk all the data labeled as hope speech to the training set of $\text{Task}_{\text{antiLGBT}}$, considered as non-anti-LGBT content. Specifically, we added $2,234$ hope speech instances.

[6]https://pypi.org/project/googletrans/

model	variation	F_1
BERT	baseline	0.94
BERT	baseline + Tamil	0.94
BERT	baseline + Hope	0.92
BERT	active learning	**0.96**

Table 3: Weighted F_1-measures on the development set for $\text{Task}_{\text{antiLGBT}}$.

Baseline augmented through hope speech data and active learning Before implementing the active learning process we added in bulk $2,234$ hope speech instances to the non-anti-LGBT content class. Then, the active learning process worked on predicting any homophobic/transphobic content within the non-hope speech instances from the pool data, i.e., from the dataset for $\text{Task}_{\text{hope}}$. From these predictions, we then integrated the top-k (with $k = 200$) instances into a newly enhanced training set and iteratively re-train and add instances until the performance stop increasing or the pool set remains empty. As a result, 194 instances have been added to the homophobic class.

4.2 $\text{Task}_{\text{hope}}$

Baseline The approach is similar to the one described for $\text{Task}_{\text{antiLGBT}}$ except that for this task we adopt a binary approach with two output units. No external data is considered in this model.

Baseline augmented through homo/transphobic data and active learning Before implementing the active learning process, we added in bulk 215 homophobic and 8 transphobic instances to the non-hope speech class. Then, we instantiated the active learning process with $k = 200$, adding 200 instances to the hope speech class.

5 Experimental Setup

No preprocessing is applied to the text, other than applying the BertTokenizer (Devlin et al., 2019). We shuffle the training set and take 10% of the data for development, preserving the class distribution through stratified random sampling (Pedregosa et al., 2011). In order to find the best hyperparameters to predict on the test set, we experimented with different batch sizes (4,8,16) for the baseline model, over an increasing number of epochs (4,6,8), testing on the development set. The combination that performed the best was a batch size of 16 over 4 epochs for both tasks, therefore we used those hyperparameters to train all models. In order to tune the network, we used the AdamW optimizer, which decouples weight decay from gradient computation, with a learning rate of 1e-5 (Loshchilov and Hutter, 2019).

model	variation	F_1
BERT	baseline	0.76
BERT	active learning	**0.77**

Table 4: Macro-averaged F_1 score for each run tested on development set.

As for the evaluation metrics, we stick to the official one: macro-averaged F_1-measure for both tasks. Since $\text{Task}_{\text{antiLGBT}}$ is a multi-class problem, we computed the weighted F_1-measure when testing on the development set.

6 Results

In this section, we present our results for both tasks. For $\text{Task}_{\text{antiLGBT}}$ we provide the results generated with the predictions of both development and test sets. For $\text{Task}_{\text{hope}}$, we present only the results on the development set.[7]

6.1 Performance on the Development Set

$\text{Task}_{\text{antiLGBT}}$ Table 3 reports the weighted F_1-measures. The best model was the active learning one, followed by the baseline and the baseline augmented with Tamil data (both 2 units less), and finally the baseline augmented with hope data (2 units less than the previous one).

$\text{Task}_{\text{hope}}$ Table 4 shows the macro-averaged F_1-measures. The highest score is obtained with the active learning approach again: F_1=0.77. The improvement over the baseline by only one unit suggests that the augmentation performed through the active learning strategy does not impact the performance significantly.

6.2 Performance on the Test Set

$\text{Task}_{\text{antiLGBT}}$ Table 5 shows the official results of our submitted runs. Contrary to the results on the development set, the baseline reached the highest score, followed by the active learning approach, the baseline augmented with Tamil data and at the end the baseline augmented with hope speech data. All the scores differ by one unit. Our baseline came fourth in the ranking. We also include macro-averaged precision and recall. The

[7] At submission time, the gold labels for the test set were not available.

model	variation	F_1	prec	rec
BERT	baseline	**0.46**	0.53	0.43
BERT	baseline + Tamil	0.43	0.49	0.41
BERT	baseline + Hope	0.42	0.45	0.41
BERT	active learning	0.44	0.49	0.41
Ablimet[(1)]		0.57	0.57	0.61
Sammaan[(2)]		0.49	0.52	0.47
Nozza[(3)]		0.48	0.58	0.45

Table 5: At the top: official macro-averaged F_1 score, precision and recall for our submissions to $\text{Task}_{\text{antiLGBT}}$ with top F_1 score highlighted. At the bottom: the performance of the top-three participants in the shared task.

relatively-low recall values indicate that the models struggle with recognizing positive instances. This result is mainly due to the nature of the dataset, which is strongly imbalanced with respect to the massive presence of instances belonging to the non-anti-LGBT class.

$\text{Task}_{\text{hope}}$ Table 6 shows the results for both submitted systems — the baseline and the baseline reinforced with the active learning approach. Both models reach the same score, positioning our team third with respect to the other participants. Once again, although the active learning approach did not impact negatively on the performance, it did not help it either.

7 Conclusions and Future Work

This paper provided a description of our participating models to the LT-EDI-ACL2022 shared tasks on hope speech detection and homophobia/transphobia detection. We addressed the two problems together, by exploiting data available in one task to create silver data for the other task.

For $\text{Task}_{\text{antiLGBT}}$, our baseline outperforms all the other reinforced approaches which make use of external data when tested on the test set. ‘For what concerns the active learning approach, it is likely that non-hope speech data do not contain homophobia or transphobia, contrary to what we expected, and therefore they do not contribute to increase the performance for $\text{Task}_{\text{antiLGBT}}$, as shown by our experiments.

For $\text{Task}_{\text{hope}}$ the active learning approach outperforms the baseline in the development set by one unit only, and it achieves the same score as the baseline in the test set, concluding that the impact of transferring data from one task to the other is not a good strategy. Nevertheless, our approaches ended up in the third and fourth position of the shared task.

model	variation	F_1	prec	rec
BERT	baseline	**0.53**	0.53	0.53
BERT	active learning	**0.53**	0.53	0.53
IIITSurat[(1)]		0.55	0.56	0.54
MUCIC[(1)]		0.55	0.54	0.55
ARGUABLY[(2)]		0.54	0.55	0.54

Table 6: At the top: Official macro-averaged F_1 score, precision and recall for our submissions to $\text{Task}_{\text{hope}}$ with top F_1 score highlighted. At the bottom: the performance of the top-three participants in the shared task.

In future work, we would like to test other transformer-based models to assess the impact of different pretraining techniques on the effectiveness of the active learning approach for these particular tasks. It would also be interesting to try different evaluation approaches for these tasks by exploring the fairness of classifiers (Dobbe et al., 2018; Mehrabi et al., 2021), with respect to minority social identities, i.e., the different members of the LGBT+ community. Specifically, we would like to investigate whether the classifiers contain unintended biases, e.g. towards specific sexual orientations, according to well-known metrics proposed to detect unfairness within toxicity detection (Borkan et al., 2019).

References

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion of The 2019 World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 491–500. ACM.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Bharathi Raja Chakravarthi. 2020. HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53, Barcelona, Spain (Online). Association for Computational Linguistics.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on Hope Speech Detection for Equality, Diversity, and Inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, John Phillip McCrae, Miguel Ángel García-Cumbreras, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Daniel García-Baena, and José Antonio García-Díaz. 2022a. Findings 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*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022b. 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.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for identification of homophobia and transophobia in multilingual youtube comments. *arXiv preprint arXiv:2109.00227*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Roel Dobbe, Sarah Dean, Thomas Krendl Gilbert, and Nitin Kohli. 2018. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics. *CoRR*, abs/1807.00553.

Paolo Rosso Elisabetta Fersini, Debora Nozza. 2020. AMI@ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4):85:1–85:30.

Hideitsu Hino. 2020. Active learning: Problem settings and recent developments. *CoRR*, abs/2012.04225.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C. Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *J. Artif. Intell. Res.*, 71:431–478.

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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35.

Arianna Muti and Alberto Barrón-Cedeño. 2020. UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo. In (Elisabetta Fersini, 2020).

Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. 2020. Hope speech detection: A computational analysis of the voice of peace. *arXiv:1909.12940*.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Oliver Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55(2):477–523.