Detection of Hate Speech using Universal Sentence Encoding and Bidirectional Long Short-Term Memory Models.

Pedro Alonso[1], György Kovács[1,*], Rajkumar Saini[1] and Marcus Liwicki[1]

[1] *Luleå University of Technology, Sweden*

Abstract

Hate speech detection is an important, but not yet resolved topic of research. The importance of the topic is particularly apparent in our age of social media. The success of old social media platforms, and emergence of new ones, allows the spread of messages (including hate speech) at an unprecedented rate. It is crucial thus to curb hate speech in order to maintain a healthy exchange of ideas and ensure that all participants feel safe expressing their views. Given the need to protect individuals from hateful language on social media, regardless of the form it takes (text, video, or audio), it is essential to find a way to foster more respectful discourse. To facilitate this, we propose a model that leverages deep learning techniques and the universal sentence encoder, aiming to navigate language nuances effectively. We applied our proposed model to the HASOC2024 task and achieved a Macro F_1-score of **0.7641**, placing us in 4^{th} position in the competition.

Keywords

Hate speech, Deep Neural Networks, Universal Sentence Encoder, Classification

1. Introduction

Hate speech continues to permeate social media exchanges [1]. The perception in recent years has been that the amount of hateful exchanges increased on **X**, (formerly known as **Twitter**). This increase, coming both from human users, as well as bots does not help any narrative[2]. Given this fact, it has become a priority to combat the hateful exchanges to create a more civilised approach to conflict resolution[3].

The mitigation of hate speech has become a critical task for maintaining healthy online environments, and protecting members of vulnerable groups. An important step towards this goal is the detection of hate speech. Automated solutions have emerged as promising, and much needed solutions for this problem, given the struggles of manual content moderation to keep pace with the sheer volume of online interactions, and the psychological toll it takes on moderators to interact with large amounts of hateful content on a daily basis [4].

1.1. Our Hybrid Approach

We have explored hybrid models combining different techniques to address some of these challenges. One such approach is our current model that combines the Universal Sentence Encoder (USE) [5, 6] with Long Short-Term Memory (LSTM) networks [7], leveraging the strengths of both techniques:

USE: It provides rich semantic representations of text, capturing contextual information effectively.

LSTM: They excel at processing sequential data, making them well suited to analyse the flow of language in potentially hateful content. This hybrid approach has shown effectiveness in detecting hate

Forum for Information Retrieval Evaluation, December 12-15, 2024, India

[*] Corresponding author.

pedro.alonso@ltu.se (P. Alonso); gyorgy.kovacs@ltu.se (G. Kovács); rajkumar.saini@ltu.se (R. Saini); marcus.liwicki@ltu.se (M. Liwicki)

speech in multiple datasets, offering improved accuracy and generalisation compared to single-technique models.

1.2. Universal Sentence encoding (USE):

The sentence encoding technique comprises several use cases which include the following [5, 6]:

- Resource constraints: USE is beneficial in scenarios with limited computational resources or where fast inference is crucial.
- Transfer learning: USE is designed for transfer learning and can be effective for a wide range of NLP tasks without extensive fine-tuning.
- General-purpose embeddings: USE provides high-quality sentence embeddings that can be used for various downstream tasks, making it versatile for general NLP applications.

2. Background

Hate speech is a complex and evolving phenomenon that has gained increased attention in the digital age. Although there is no universally accepted, precise and productive definition, most would agree that hate speech means public speech that expresses hate, encourages violence or promotes prejudice against individuals or groups based on attributes such as race, religion, ethnicity, sexual orientation, or gender identity.

2.1. Definition of Hate Speech

The lack of a consistent definition poses challenges for researchers and policymakers. Various institutions and scholars have proposed their own definitions:

- The Cambridge Dictionary defines hate speech as "public speech that expresses hate or encourages violence towards a person or group based on something such as race, religion, sex, or sexual orientation" [8].
- Facebook defines hate speech as "a direct attack on people based on protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability" [9].
- Fortuna et al. describe hate speech as "language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other" [3].

These definitions highlight the complexity of categorising hate speech and the need for a nuanced approach to its detection and mitigation in two ways. For one, these definitions highlight the complexity of the problem by their differences. In the definition of the Cambridge Dictionary, the expression of hate is sufficient to qualify content as hateful. While in the description of Fortuna et al. language needs to attack, diminish, or incite violence/hate. And according to facebook the message has to be a direct attack. Moreove, the complexity arises from the fact that each definition contains expressions that themselves may need to be defined. For example, what constitutes as a direct attack? What would encourage someone to violence? This would surely be different from person to person, and consequently, it is likely that different annotators would judge this question differently, when having to annotate for the presence or absence of hate speech.

2.2. Prevalence and Consequences of Hate Speech in Online Platforms

The proliferation of social media and online communication has increased the visibility and spread of hate speech. Studies have shown that the prevalence of hate speech on online platforms can range from 30% to 40% of user-generated content, depending on the platform and definition used [10, 11, 12]. The consequences of online hate speech are far-reaching and severe:

- Psychological harm: Victims of hate speech often experience stress, anxiety, and depression [13].
- Social cohesion erosion: Hate speech can contribute to the marginalisation of minority groups and increase societal divisions.
- Potential for real-world violence: Online hate speech has been linked to an increase in hate crimes and can be a precursor to more severe forms of discrimination [14].
- Threat to democratic discourse: Hate speech can stifle free expression by intimidating certain groups from participating in online discussions [3].

The challenge of addressing hate speech while preserving freedom of expression has led to varying approaches in different countries. Although many developed democracies have implemented laws restricting hate speech, the United States has consistently protected it under the First Amendment, creating a complex landscape for global online platforms [15].

Practical strategies for detecting and mitigating hate speech become increasingly crucial as the digital sphere evolves. This background underscores the importance of developing sophisticated, context-aware systems to identify and address hate speech in online environments.

2.3. Importance and Challenges of Hate Speech Detection

Hate speech detection is important for several reasons, some of which are:

- Protecting vulnerable groups: It helps safeguard individuals and communities from targeted harassment and discrimination.
- Maintaining online safety: It creates a safer and more inclusive digital environment.
- Preventing offline consequences: Early detection can help prevent the escalation of online hate into real-world violence or discrimination.

However, the task of hate speech detection faces several challenges:

- Contextual nuances: Understanding the context and intent behind potentially offensive language is complex.
- Evolving language: Hate speech often adapts and uses coded language to evade detection.
- Balancing free speech: There is a need to balance removing harmful content and preserving freedom of expression.

3. Methods

Our proposed model architecture uses a hybrid approach, combining the Universal Sentence Encoder (USE) with a Bidirectional Long-Short-Term Memory (BiLSTM) network. This architecture aims to capture contextual embeddings (via USE), and sequential information (via LSTM) from the input text. First, we describe the datasets used and then our hybrid approach.

3.1. Datasets

Our experiments utilise a dataset comprising four distinct hate speech corpora. HASOC 2019, 2020, 2021 [16, 17, 18] and the offensive language identification dataset (OLID) [19]. This approach allows for a comprehensive analysis of hate speech in multiple years and contexts, which can help us detect new and unforeseen hate discourse.

3.1.1. Combined Dataset Description

The HASOC (Hate Speech and Offensive Content Identification) datasets from 2019, 2020, and 2021 [16, 17, 18] provide a rich source of multilingual hate speech data. Each year's dataset offers unique characteristics:

Dataset	Description
HASOC 2019 [16]	Contains 5,852 English tweets.
HASOC 2020 [17]	Contains 5,335 English tweets.
HASOC 2021 [18]	Contains 5,484 English tweets.
OLID [19]	Contains 14,100 English tweets.

Table 1
Summary of HASOC and OLID datasets

3.1.2. Data Pre-processing

Our pre-processing pipeline includes the following:

- Removal of emojis, emoticons and special characters
- Lowercasing and tokenisation
- Replacing of URLs, user mentions and hashtags by placeholders such as URL, USER and HASHTAG.

These methods helped us to have a cleaner text that is more agreeable to further processing.

3.1.3. 5-Fold Cross-Validation Approach and Final Submission Processing

To ensure reliable model evaluation, we implement a 5-fold cross-validation strategy. This evaluation format involves:

- Splitting the combined dataset into five equal parts
- Training the model on four parts and testing on the remaining part
- Testing on the test set 5 times to cover all data
- Averaging the performance metrics across all folds

Lastly, before submitting the final results, we conducted majority voting on three different runs of our results, which helped us further filter down the test results.

3.2. Model description: Universal Sentence Encoder

The model begins with the Universal Sentence Encoder (USE), a pre-trained model that encodes text into high-dimensional vectors [5]. We utilise the USE (version 4) as a fixed feature extractor, setting its layers as non-trainable to benefit from its pre-trained knowledge while reducing computational costs.

3.2.1. Model description: Bidirectional LSTM Layers

The USE output is reshaped and fed into two stacked Bidirectional LSTM layers. The first BiLSTM layer contains 1500 units and returns sequences, while the second contains 1000 units. These layers process the input in both forward and backward directions, allowing the model to capture complex temporal dependencies and contextual information from past and future states.

3.2.2. Model description: Dense Layers and Regularisation

Following the BiLSTM layers, the model incorporates dense layers of units 1024, 512, 300, and 256, each with a ReLU activation function. To combat overfitting, we utilised several regularisation techniques:

- Each dense layer has an L2 or weight decay, regularizator.
- Constrained max norm, to 3, on the kernel weights of the dense layers.
- The dropout layers have rates of 0.5 and 0.4 after each dense layer.

3.2.3. Model description: Output Layer

The final layer is a single-unit dense layer with a sigmoid activation function to get a probability score for the classification task.

3.2.4. Model description: Model Compilation

We used the Adam Optimiser with a learning rate of 1e-3 and binary cross-entropy as the loss function, which is common in binary classification tasks. We also implement gradient clipping (clipnorm=1.0) to prevent the gradient from exploding. The process is visualized in fig. 1

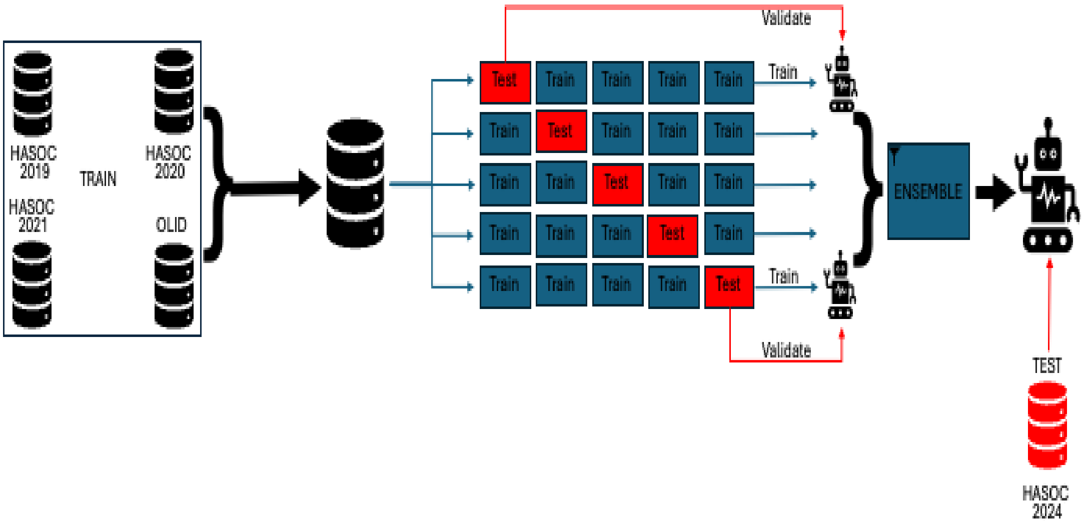

Figure 1: Visual representation of the methodology.

3.3. Performance Metrics and Cross-Validation

We evaluated our model using accuracy as the primary metric for selecting the best model. To ensure a robust performance assessment, we implement a five-fold cross-validation approach. This evaluation involves training and evaluating the model on five different train-validation splits of the dataset, then averaging the results to obtain a more reliable estimate of the model's performance.

3.4. Model Ensemble

We create an ensemble of models trained on different folds to further improve prediction stability and accuracy. The final prediction is obtained by averaging the output of these individual models.

This hybrid USE-BiLSTM architecture, combined with regularisation techniques and ensemble learning, aims to effectively capture the nuances of hate speech while maintaining generalisation capability.

Layer (type)	Output Shape	Param #
Lambda (Lambda)	?	0
reshape (Reshape)	(None, 1, 512)	0
bidirectional (Bidirectional)	(None, 1, 3000)	24,156,000
bidirectional_1 (Bidirectional)	(None, 2000)	32,008,000
dense (Dense)	(None, 1024)	2,049,024
dropout (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 512)	524,800
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 300)	153,900
dropout_2 (Dropout)	(None, 300)	0
dense_3 (Dense)	(None, 256)	77,056
dropout_3 (Dropout)	(None, 256)	0
dense_4 (Dense)	(None, 1)	257

Table 2
Deep Neural Network Architecture

- **Total params**: 176,907,113 (674.85 MB)
- **Trainable params**: 58,969,037 (224.95 MB)
- **Non-trainable params**: 0 (0.00 B)
- **Optimizer params**: 117,938,076 (449.90 MB)

4. Results and Discussion

This paper focused on subtask A of the HASOC 2024 competition. Our results are only for the English part. The submission comprised three different runs of the model's code restarting after each completion, in which the previously mentioned step was done 3.1.3. In Table 3, we have the results supplied to us by the organisers; here, we see that our model got a 4^{th} place with a Macro F_1-score of **0.7641**.

Table 3
Team Rankings by Best Macro F1 Score

Rank	Team Name	Best Macro F1 Score
1	FiRC-NLP	0.8133
2	Jing	0.8005
3	wangkongqiang	0.7939
4	**TheNorthAtLTU**	**0.7641**
5	TextTitans	0.7561
6	Abishikta	0.7104
7	MUCS	0.5653
8	KK_IIITR_Research_Lab	0.5492

As the table 3 shows, our score did not vary more than 0.05 from the top groups. Considering that we did not actively use a transformer model variation, the missing points can be seen as a trade-off between performance and energy consumption.

We also analysed our results, considering the text lengths, which could guide us on why or where the errors could be.

In fig. 2, we show the distribution we obtained for one of our experiments of the predicted classes. We also plot a histogram of text length that we obtain from our predictions set to look for some insight that we may have overlooked that the text distributions could provide; we show this in fig. 3.

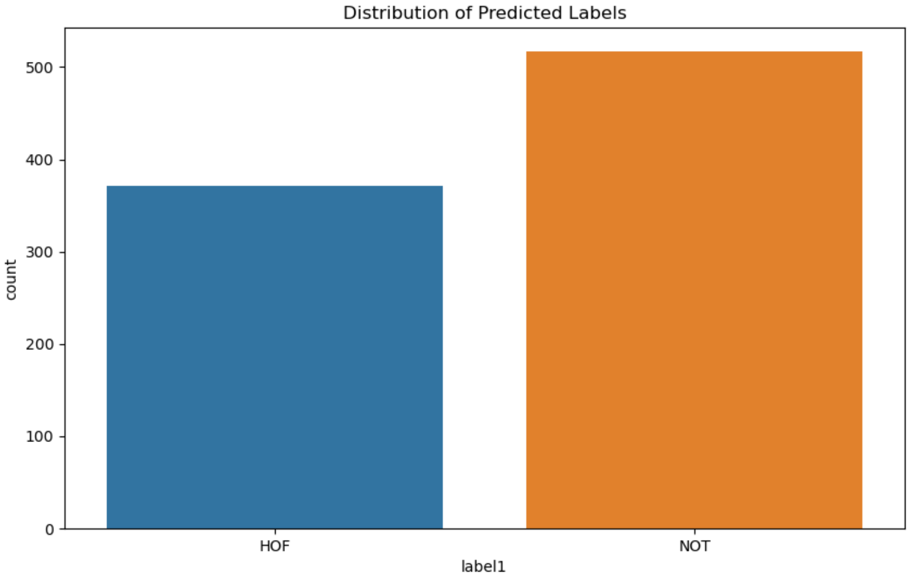

Figure 2: Predicted class distribution.

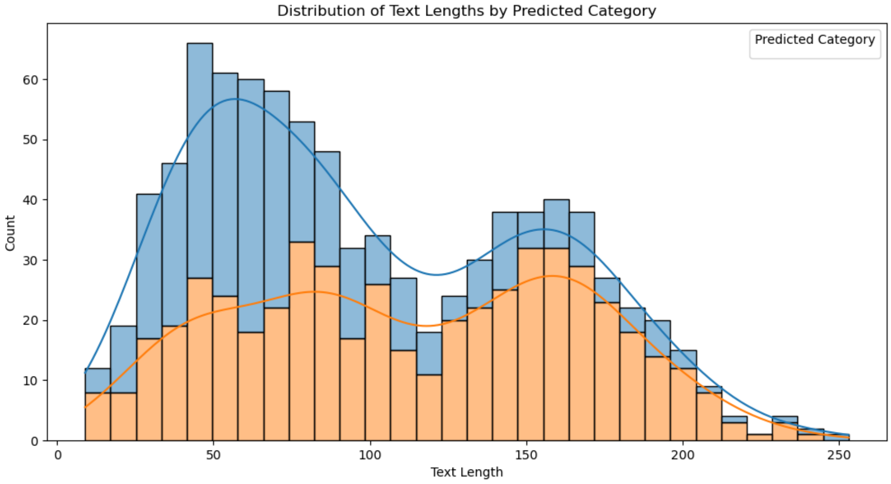

Figure 3: Histogram of text length per class.

5. Conclusion

Hate speech detection remains a significant and unsolved problem in the field of natural language processing (NLP) due to the natural evolution of language. While it may continue to be unsolved in the near future, our approach aims to introduce a consistent method for detecting hate speech.

Our approach, which leverages hate speech datasets from the HASOC competitions and the OLID dataset, has demonstrated its effectiveness. This combination allowed us to achieve the 4th position with a Macro F_1-score of **0.7641**. This achievement serves as evidence that combating online hate speech can be accomplished without relying heavily on resource-intensive models like transformers. Our approach has the potential to make a significant impact on this important societal issue.

6. Future Work

We want to incorporate a more comprehensive range of data for classification, which could help further refinement. We will also explore new ideas for improving model efficiency and reducing power consumption and will strive to incorporate multilingual capabilities.

We hope that our research has put forward a new direction for effective hate speech detection, but there are several avenues for future exploration and improvement; some of the future directions include:

6.1. Incorporating a Wider Array of Data for Classification

A key focus of our future work will be expanding our dataset to cover a broader range of hate speech types and contexts. By incorporating data from various sources, languages, and cultural contexts, we aim to develop a more robust and generalisable model [20]. This expanded data set will help address the current limitations of dataset-specific classifiers and move toward a more universal hate speech detection system [21].

6.2. Improving Model Efficiency and Reducing Power Consumption

As we work towards more sophisticated models, we will also focus on optimising their efficiency:

- Exploring model compression techniques to reduce the computational requirements of our hate speech detection systems[22].
- Investigating energy-efficient architectures that can maintain high performance while minimising power consumption[23, 24].
- Developing lighter models suitable for edge computing, enabling real-time hate speech detection on user devices utilising frameworks like hyperdimensional computing[25].

By following these research directions, we hope to develop a next-generation hate speech detection system that is more accurate, efficient, and adaptable to the current and future hate speech content created by the online world.

Declaration on Generative AI

During the preparation of this work, the author(s) used Writefull and Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] B. Vidgen, L. Derczynski, Challenges and frontiers in abusive content detection, Proceedings of the Third Workshop on Abusive Language Online (2019) 80–93.

[2] Y. Jiang, X. Xiang, H. Yin, Hate speech detection on twitter: A comprehensive review, ACM Computing Surveys 55 (2023) 1–38.

[3] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[4] R. Spence, A. Bifulco, P. Bradbury, E. Martellozzo, J. DeMarco, The psychological impacts of content moderation on content moderators: A qualitative study, Cyberpsychology: Journal of Psychosocial Research on Cyberspace 17 (2023) Article 8. URL: https://cyberpsychology.eu/article/view/33166. doi:10.5817/CP2023-4-8.

[5] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. St. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, B. Strope, R. Kurzweil, Universal sentence encoder for English, in: E. Blanco, W. Lu (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing:

System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 169–174. URL: https://aclanthology.org/D18-2029. doi:10.18653/v1/D18-2029.

[6] S. Sarkar, D. Feng, S. K. K. Santu, Exploring universal sentence encoders for zero-shot text classification, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2022, pp. 135–147.

[7] J. Nowak, A. Taspinar, R. Scherer, Lstm recurrent neural networks for short text and sentiment classification, in: Artificial Intelligence and Soft Computing: 16th International Conference, ICAISC 2017, Zakopane, Poland, June 11-15, 2017, Proceedings, Part II 16, Springer, 2017, pp. 553–562.

[8] D. Crystal, The Cambridge encyclopedia of the English language, Cambridge university press, 2018.

[9] M. Hietanen, J. Eddebo, Towards a definition of hate speech—with a focus on online contexts, Journal of Communication Inquiry 47 (2023) 440–458.

[10] M. A. Chekol, M. A. Moges, B. A. Nigatu, Social media hate speech in the walk of ethiopian political reform: analysis of hate speech prevalence, severity, and natures, Information, Communication & Society 26 (2023) 218–237.

[11] S. A. Castaño-Pulgarín, N. Suárez-Betancur, L. M. T. Vega, H. M. H. López, Internet, social media and online hate speech. systematic review, Aggression and violent behavior 58 (2021) 101608.

[12] J. Kansok-Dusche, C. Ballaschk, N. Krause, A. Zeißig, L. Seemann-Herz, S. Wachs, L. Bilz, A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena, Trauma, violence, & abuse 24 (2023) 2598–2615.

[13] Wikipedia, Hate speech, https://en.wikipedia.org/wiki/Hate_speech, 2024. Accessed: 2024-09-06.

[14] P. Saha, B. Mathew, P. Goyal, A. Mukherjee, Hate speech review in the context of online social networks, Aggression and Violent Behavior 48 (2019) 108–118.

[15] American Library Association, Hate speech and hate crime, https://www.ala.org/advocacy/intfreedom/hate, 2024. Accessed: 2024-09-06.

[16] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th forum for information retrieval evaluation, 2019, pp. 14–17.

[17] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, 2020, pp. 29–32.

[18] S. Modha, T. Mandl, G. K. Shahi, H. Madhu, S. Satapara, T. Ranasinghe, M. Zampieri, Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages and conversational hate speech, in: FIRE (Working Notes), 2021.

[19] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, 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, Volume 1 (Long and Short Papers), 2019, pp. 1415–1420.

[20] S. MacAvaney, H.-R. Yao, E. Yang, K. Russell, N. Goharian, O. Frieder, Hate speech detection: Challenges and solutions, PloS one 14 (2019) e0221152.

[21] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, A. Hussain, Hate speech detection: A comprehensive review of recent works, Expert Systems (2024) e13562.

[22] Y. Zhang, S. Gao, H. Huang, Exploration and estimation for model compression, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 487–496.

[23] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in nlp, arXiv preprint arXiv:1906.02243 (2019).

[24] C. Zonios, V. Tenentes, Energy efficient speech command recognition for private smart home iot applications, in: 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST), IEEE, 2021, pp. 1–4.

[25] P. R. Genssler, A. Vas, H. Amrouch, Brain-inspired hyperdimensional computing: How thermal-

friendly for edge computing?, IEEE Embedded Systems Letters 15 (2022) 29–32.