Addressing Hate Speech: ATLANTIS for Efficient Hate Span Detection

Niyar R Barman[1,2], Krish Sharma[1,2], Yashraj Poddar[1], Advaitha Vetagiri[1,3] and Partha Pakray[1]

[1]National Institute of Technology, Silchar, Assam, India - 788010
[2]Both authors contributed equally to this research
[3]Corresponding author.

Abstract

Hate speech poses significant challenges to maintaining healthy online conversations, and automated systems are crucial for its accurate detection and mitigation. In this paper, we (CNLP-NITS-PP) introduce ATLANTIS (Attentive Transformer-LSTM for Named Entity and Token Identification System), a robust model designed to address the pervasive issue of hate speech in online social media platforms. ATLANTIS focuses on hate span identification within sentences labeled as hate speech, framed as a sequence labeling task using BIO notation. Leveraging a Hate dataset enriched with Named Entity Recognition (NER) tags, ATLANTIS effectively identifies hate speech spans within the text by combining contextualized representations and sequential modeling. The empirical results showcase ATLANTIS's effectiveness in isolating explicit signs of hate from a contextual backdrop, offering a promising solution for creating safer online environments. We achieve a macro F1 score of 0.488 on the public test set and 0.508 on the private test set. This work not only lays the foundation for future advancements in hate-span detection but also emphasizes the importance of model efficiency, interpretability, and expanded training data that encompass diverse linguistic nuances and evolving hate speech trends. Code is available at https://github.com/niyarrbarman/hasoc23

Keywords

Hate Speech Detection, Named Entity Recognition (NER), Sequence Labeling, Natural Language Processing, Transformer, BiLSTM

1. Introduction

Social media platforms like Twitter and Facebook have become commonplace in modern life, giving people worldwide easy access to voice their thoughts and connect. However, the open nature of these platforms also allows harmful content like hate speech, harassment, and threats aimed at vulnerable groups to spread [1]. This has created an urgent need for automated systems that accurately recognise abusive language to maintain healthy online conversations [2].

Forum for Information Retrieval Evaluation, December 15–18, 2023, Goa, India
barmanniyar@gmail.com (N. R. Barman); iamkrish9090@gmail.com (K. Sharma); yash.raj.poddar.yp@gmail.com (Y. Poddar); advaitha21_rs@cse.nits.ac.in (A. Vetagiri); partha@cse.nits.ac.in (P. Pakray)
https://niyarrbarman.github.io/ (N. R. Barman)
0009-0001-2112-2491 (N. R. Barman); 0009-0007-7001-7480 (K. Sharma); 0009-0007-6119-3255 (Y. Poddar); 0000-0002-0651-4171 (A. Vetagiri); 0000-0003-3834-5154 (P. Pakray)

A significant hurdle is that offensive content can take many linguistic forms, necessitating context-aware models to pinpoint the specific snippets of text that render a post hateful or abusive [3]. Furthermore, implicit forms of hate speech, like veiled insults, require deducing pragmatic implications [4] rather than just spotting explicit derogatory terms [5]. This has driven recent research into models for singling out spans of text that communicate hateful intent within a given post [6].

This paper tackles the problem of hate span identification within sentences labelled as hate speech in the HASOC 2023 [7] shared task [8]. In this paper, we delve into the challenges and innovations of the HASOC subtrack at FIRE 2023, focusing on the 'Detection of Hate Spans and Conversational Hate-Speech,' as outlined by Satapara et. al [9]. Given an English social media sentence already deemed hateful, the goal is to pinpoint contiguous spans of tokens that relay its hateful purpose. This is framed as a sequence labelling task using BIO notation, where each token is tagged as the Beginning (B), Inside (I), or Outside (O) of a hate span [10].

The HASOC dataset provides ground truth BIO tag sequences for abusive sentences from public hate speech sources [8]. Participants construct models to predict these spans in test sentences without extra preprocessing to avoid incongruities. This focused evaluation enables the systematic development of context-aware models and techniques for fine-grained hate speech analysis, moving beyond the binary classification of posts [11].

We present our proposed model design and tactic for the hate span identification task, harnessing contextualised representations and sequential modelling [12]. Results showcase our techniques' efficacy in isolating explicit signs of hate from a contextual backdrop. By classifying specific linguistic cues and semantic relationships that encode hate, our method provides insights into the underlying fabric of abusive language [6].

2. Application and Target Audience

The research presented in this paper holds significant promise in tackling the pervasive problem of hate speech on online social media platforms. ATLANTIS, the hate span detection system that has been developed, carries practical implications for content moderation, user safety, and the improvement of online discussions. By precisely identifying and extracting hate spans from hateful sentences, ATLANTIS equips social media platforms to more efficiently filter and eliminate hateful content, thereby promoting a safer and more inclusive online environment. Furthermore, this technology can serve as a valuable tool for gaining insights into the prevalence and dynamics of hate speech, assisting researchers and policymakers in formulating evidence-based strategies to combat online hatred.

This research paper is intended for a diverse audience encompassing various stakeholders concerned with the detection and mitigation of hate speech. Content moderators and social media platform administrators will find valuable insights and methodologies within as they work towards maintaining respectful and secure online communities. Researchers in the fields of natural language processing (NLP) and machine learning will appreciate the detailed methodology and architecture of the ATLANTIS model, which represents an advancement in state-of-the-art hate span detection. Policymakers and organizations focused on addressing online hate speech will also gain valuable insights into the potential of machine learning-based

solutions for addressing this pressing issue. Furthermore, educators and students studying NLP, machine learning, and technology ethics can utilize this paper as a resource for understanding the development and application of advanced models for hate speech detection. Ultimately, this research paper aims to engage a broad and diverse audience, fostering collaboration and innovation in the ongoing effort to create safer online spaces.

3. Objective

The primary objective of this research is to create a hate span detection system capable of pinpointing and extracting uninterrupted sequences of tokens found within hateful sentences, which we refer to as "hate spans". These hate spans are characterized as consecutive sets of tokens within a sentence that collectively expresses explicit hatefulness. The aim of this shared task is to automatically identify and extract all such hateful spans from preprocessed sentences. The hate span detection task is approached as a sequence labeling problem, wherein each token in a sentence is labeled with a specific tag to indicate its association with a hateful span. The labeling follows the BIO notation, with 'B' signifying the beginning of a hate span, 'I' denoting the continuation of a hate span, and 'O' indicating all other tokens that are not part of any hate span within the sentence.

The goal is to develop a machine-learning model to accurately predict the correct sequence of BIO tags for each token in a given sentence, effectively detecting and delineating hate spans within the text.

4. Proposed Methodology

The methodology employed to address the issue of hate speech at scale through the ATLANTIS model comprises a systematic approach encompassing data preprocessing, tokenization, model architecture, and the classification process. Leveraging the HateNorm23 dataset, which features text samples paired with Named Entity Recognition (NER) tags categorizing each word as 'B' (signifying the start of a hate span), 'I' (indicating inclusion within a hate span), or 'O' (denoting other), we conduct word-level tokenization to segment the text into meaningful units. A custom tokenizer is then fine-tuned on the dataset to tailor tokenization for hate span detection. The ATLANTIS model adopts a multi-stage architecture, initially processing tokenized text through a custom transformer section followed by a bidirectional long short-term memory (Bi-LSTM) [13] section. The transformer captures contextual information and relationships, while the Bi-LSTM captures sequential dependencies. Subsequently, fused representations from these sections traverse fully connected layers for the conclusive classification task. Detailed insights into the architecture, hyperparameters, and experimental findings will be presented to substantiate ATLANTIS's efficacy in mitigating hate speech at scale.

ATLANTIS consists of three primary components:

Transformer Encoder Block: The Transformer [14] block is a foundational component for capturing contextual relationships within sequences. Its self-attention mechanism enables the model to weigh the significance of each word in relation to others, allowing it to understand

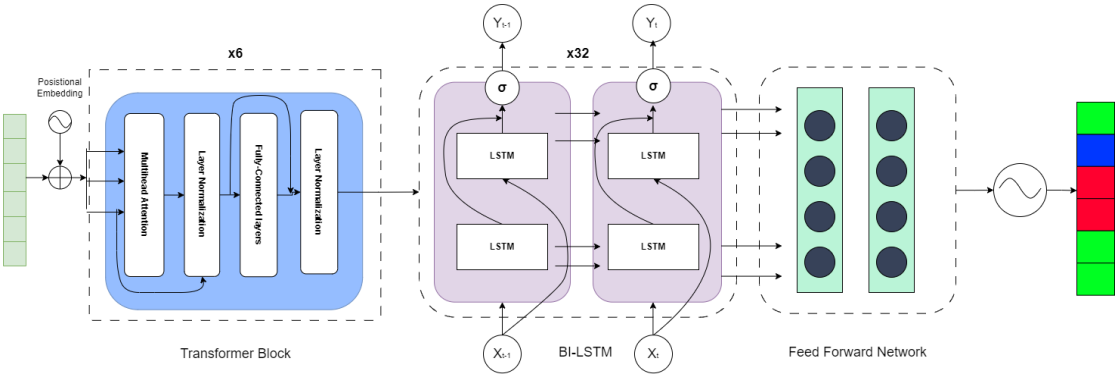


Figure 1: Architecture of ATLANTIS, comprising three primary components — Transformer Encoder Block, BiLSTM Layer, and Sequential Block with FC Layers—designed for effective sequence understanding and Hate Span Identification

complex dependencies and semantic connections. This block excels at learning hierarchical features from the input data, providing a solid basis for understanding the underlying patterns in the sequential data, which is particularly crucial in NLP tasks.

BiLSTM Layer: The BiLSTM layer complements the Transformer's strengths by effectively capturing sequential dependencies in the data. By incorporating a BiLSTM layer, the model can capture fine-grained temporal relationships and contextual nuances that might be missed by the Transformer alone. This is especially valuable for NER, where identifying entities often relies on sequential patterns.

Sequential Block with FC Layers: The Sequential Block, containing Fully Connected layers, serves as a vital element for transforming the enriched features from the preceding blocks into a suitable format for making predictions. These FC layers allow for nonlinear transformations and higher-level abstractions, enabling the model to learn complex mappings from the learned representations to the target NER labels.

Engineering Decisions: We aimed to identify a solution that excels in performance and efficiency. Our approach led us to employ a sequence of six transformer blocks. Upon extending the number of blocks, we observed a period during which the F1 score plateaued, roughly around 9 to 10 blocks. Subsequently, the score rapidly declined, indicative of overfitting taking hold.

Regarding the BiLSTM layers, we integrated a single BiLSTM layer for the ultimate modeling phase. Elevating the count of BiLSTM layers increased the model's complexity, rendering it more challenging to train and subsequently slowing down inference processes.

We settled on a configuration of `num_heads = 4` for the transformer block. Introducing additional `num_heads` led to a stage of diminishing returns. Given the limited size of our dataset, the model tended to memorize the training data rather than exhibiting the capacity to generalize to novel data. This phenomenon, in turn, resulted in overfitting or diminished performance.

Adam was used as the optimizer with `learning_rate = 1e-3`

5. Dataset

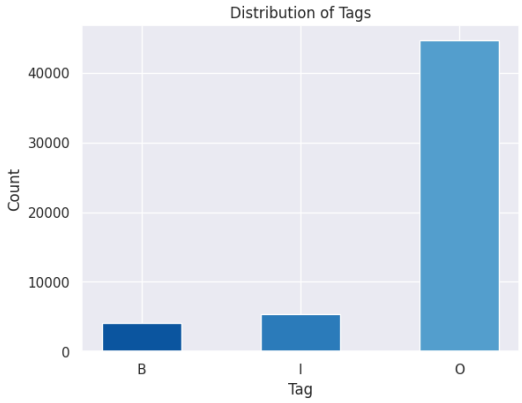


Figure 2: Visualization of Token-level BIO Tags Distribution in the dataset

The dataset [15] comprises a total of 2421 data points. We partitioned this dataset into an 80:10:10 ratio, allocating segments for training, validation, and testing purposes. Within the dataset, a sum of 8165 distinct words can be found. The visualization of the dataset is presented in Figure 2. Notably, hate speech constitutes 17.422% of the entire dataset.

6. Results and Analyses

In this section, we present the results of our experiments, organized into three subsections: Baseline Methods, Intrinsic Results, and Extrinsic Results. We discuss the models we used in the Baseline Methods section and provide details on the intrinsic and extrinsic performance of our approach.

6.1. Baseline Methods

To establish a benchmark for our experiments and assess the effectiveness of our proposed method, we employed the following baseline models:
Pretrained BERT: BERT [12] has shown remarkable success in various natural language processing tasks, and we included it as a reference to evaluate the performance of our approach against a state-of-the-art model.
Transformer Encoder: The incorporation of the Transformer [14] Encoder, in our study serves a dual purpose. Firstly, it provides a reference point for evaluating the performance of our approach. Secondly, it underscores the effectiveness of the encoder layers, equipped with self-attention mechanisms, which play a key role in the remarkable success of BERT and similar models across various natural language processing tasks.
BiLSTM: BiLSTM [13] networks have been widely used for sequence labeling tasks, and we included this baseline to evaluate our approach against a more traditional sequence labeling model.

Table 1
Performance metrics of baseline models for different tags

Model	Precision	Recall	F1-Score
BERT	0.58	0.56	0.57
Transformer	0.82	0.79	0.80
Bi-LSTM	0.56	0.61	0.58

6.2. Intrinsic Results

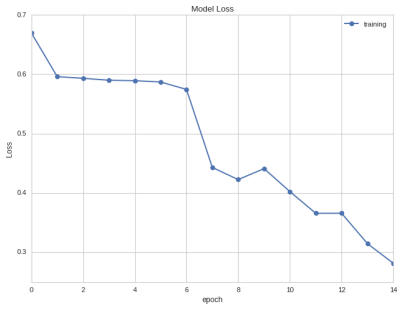
In this subsection, we present the intrinsic results of our approach to the validation set. We discuss the performance of our model and provide a detailed analysis of the results.

Our model's performance on the validation set was evaluated using various metrics, including precision, recall and F1-score. They have been presented in Table 2.

Table 2
ATLANTIS performance metrics for different tags

BIO-Tags	Precision	Recall	F1-Score
B	0.83	0.78	0.77
I	0.72	0.81	0.76
O	0.97	0.95	0.96

Figure 3: Model's F1 score variation over epochs



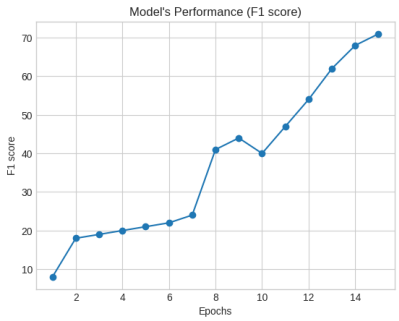


Figure 4: Model's Loss variation over epochs

The graph in Figure 4 illustrates the model's loss convergence during training. As we can observe, the loss steadily decreases over epochs, indicating that our model effectively learns to minimize the prediction errors.

Figure 3 showcases the improvement in the F1-score over training epochs. The upward trend in F1-score suggests that our model becomes increasingly proficient at correctly identifying and labeling entities in the validation data as training progresses.

6.3. Extrinsic Results

Table 3
Performance metrics of baseline models for different tags

Model	Macro F1-Score	
	Public Test Set	**Private Test Set**
BERT	0.303	0.360
Transformer	0.446	0.473
Bi-LSTM	0.315	0.324
ATLANTIS	**0.488**	**0.508**

In this subsection, we present the extrinsic results of our approach to the competition test set. We report public and private test scores, commonly used in Kaggle competitions to evaluate model performance on unseen data. Table 3 summarizes our model's public and private test scores and compares them with the baseline models.

7. Related Work

In this section, we review several relevant studies that contribute to the understanding and development of hate speech detection, offensive language detection, and related natural language processing tasks. These works collectively provide insights into various approaches and techniques employed in this field.

In Qian et al.'s (2019)[16] study [14], a new challenge called generative hate speech intervention was introduced. The authors augmented their research with two comprehensive datasets obtained from Reddit and Gab, which contained intervention responses collected from crowdsourcing. The assessment of three generative models, specifically Seq2Seq, VAE, and RL, revealed areas where hate speech intervention methods could be enhanced.

In the work conducted by Alshalan et al. [17], they tackled the problem of hate speech in the Arabic Twittersphere. They introduced a dataset consisting of 9316 tweets categorized into hate speech, abuse, and normalcy. Their assessment encompassed various models, including CNN, GRU, CNN + GRU, and BERT. Among these models, CNN emerged as the most effective, achieving superior performance with an F1-score of 0.79 and an AUROC of 0.89.

In the research conducted by Elalami et al. [18], they introduced a transfer learning strategy for detecting offensive language in multiple languages. This approach leveraged several BERT models, such as BERT, mBERT, and AraBERT. Their results were outstanding, surpassing the performance of current leading methods that employ joint-multilingual and translation-based approaches. This study underscored the robustness of BERT models in the context of Multilingual Offensive Language Detection.

Ozler et al. [19] explored the application of BERT for multi-label and multi-domain incivility detection tasks. They successfully established a new state-of-the-art performance across various datasets. The study suggested that direct data combination from multiple domains yielded superior results compared to more intricate training methods.

The study by Hoang et al. [20] introduced ViHOS, a novel Vietnamese dataset for hate and offensive span detection, containing 26,467 annotated spans in 11,056 comments. Baseline models, including XLM-RBase, XLM-RLarge, PhoBERTBase, and PhoBERTLarge, were evaluated, with the XLM-RLarge model leading with an F1-score of 0.7770. The study found that detecting multiple spans outperformed single-span detection in Vietnamese hate speech.

Lample et al. [10] introduced a discriminative parsing-based approach for nested named entity recognition, demonstrating strong performance on top-level and nested entities. However, the study acknowledged a limitation in terms of speed compared to conventional flat techniques. The paper advocated for reconsidering the exclusion of embedded entities in NER corpora, highlighting the substantial information loss incurred by this design choice.

Ma (2016) [21] presented a neural network architecture for sequence labeling, representing an end-to-end model without needing task-specific resources, feature engineering, or data preprocessing. The study attained state-of-the-art performance on two linguistic sequence labeling tasks, outperforming prior state-of-the-art systems.

Peters et al. (2017) [22] proposed a simple semi-supervised approach using pre-trained neural language models to enhance token representations in sequence tagging models. Their approach consistently outperformed state-of-the-art models in NER and Chunking datasets. Notably, the study showed that including both forward and backward language models consistently improved performance.

These related works collectively contribute valuable insights and methodologies that inform the development of hate speech detection and associated natural language processing tasks, showcasing the advancements and challenges in this field.

8. Conclusion and Future Scope

In this research, we have presented ATLANTIS (Attentive Transformer-LSTM for Named Entity and Token Identification System), a robust model designed to combat hate speech at scale. Leveraging a Hate dataset with detailed Named Entity Recognition (NER) tags, ATLANTIS effectively identifies hate speech spans within textual content. Our multi-stage architecture, comprising a custom transformer and bidirectional LSTM, captures contextual information and sequential dependencies, facilitating precise hate span classification. Empirical results demonstrate ATLANTIS's effectiveness in this critical task. As we continue to address the pressing issue of hate speech in digital spaces, ATLANTIS offers a promising solution for safer online environments.

The work presented here lays the foundation for future advancements in hate span detection. Further improvements in model efficiency and interpretability, along with expanded training data encompassing diverse linguistic nuances and evolving hate speech trends, hold promise. Investigating the integration of real-time monitoring and incorporating user-specific context may enhance the model's capabilities in dynamically changing online environments. Additionally, exploring multilingual and cross-platform hate speech detection is vital for broader impact. As technology evolves, ATLANTIS and its successors are poised to play a pivotal role in fostering safer, more inclusive digital spaces.

Acknowledgments

We wish to extend our appreciation to the Computer Science and Engineering Department of the National Institute of Technology Silchar for granting us the opportunity to carry out our research and experiments. We are grateful for the support, resources, and research environment offered by the CNLP & AI Lab at NIT Silchar.

References

[1] B. Vidgen, L. Derczynski, (2020), Directions in abusive language training data, a systematic review: Garbage in, garbage out. PloS one 15 (2020).

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

[3] A. Vetagiri, P. K. Adhikary, P. Pakray, A. Das, "CNLP-NITS at SemEval-2023 Task 10: Online sexism prediction, PREDHATE!", In the 17th International Workshop on Semantic Evaluation SemEval 2023 Toronto, Canada July 9-14, 2023.

[4] D. Jurgens, L. Hemphill, E. Chandrasekharan, A just and comprehensive strategy for using NLP to address online abuse, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[5] A. Vetagiri, P. K. Adhikary, P. Pakray, A. Das, "Leveraging GPT-2 for Automated Classification of Online Sexist Content", In Exist 2023 Lab at CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece, 2023.

[6] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, (2021), Hatexplain: A benchmark dataset for explainable hate speech detection. Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 14867–14875.

[7] S. Masud, M. A. Khan, M. S. Akhtar, T. Chakraborty, Overview of the HASOC Subtrack at FIRE 2023: Identification of Tokens Contributing to Explicit Hate in English by Span Detection, in: Working Notes of FIRE 2023 - Forum for Information Retrieval Evaluation, CEUR, 2023.

[8] 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, Proceedings of the 11th Forum for Information Retrieval Evaluation, 2019.

[9] S. Satapara, S. Masud, H. Madhu, M. A. Khan, M. S. Akhtar, T. Chakraborty, S. Modha, T. Mandl, Overview of the HASOC subtracks at FIRE 2023: Detection of hate spans and conversational hate-speech, in: Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation, FIRE 2023, Goa, India. December 15-18, 2023, ACM, 2023.

[10] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603 (2016) 01360.

[11] Z. Zhang, L. Luo, Hate speech detection: A solved problem? The challenging case of long tail on Twitter, Semantic Web 10 (2019) 925–945.

[12] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810 (2018) 04805.

[13] M. Schuster, K. Paliwal, Bidirectional recurrent neural networks, Signal Processing, IEEE Transactions on 45 (1997) 2673 – 2681. doi:10.1109/78.650093.

[14] A. Vaswani, Attention is all you need, 2017. URL: https://arxiv.org/abs/1706.03762.

[15] S. Masud, M. Bedi, M. A. Khan, M. S. Akhtar, T. Chakraborty, Proactively reducing the hate intensity of online posts via hate speech normalization, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3524–3534. URL: https://doi.org/10.1145/3534678.3539161. doi:10.1145/3534678.3539161.

[16] J. Qian, A. Bethke, Y. Liu, E. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, 2019. URL: https://arxiv.org/abs/1909.04251v1.

[17] R. Alshalan, H. Al-Khalifa, A deep learning approach for automatic hate speech detection in the saudi twittersphere, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/23/8614. doi:10.3390/app10238614.

[18] F. zahra El-Alami, S. Ouatik El Alaoui, N. En Nahnahi, A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model, Journal of King Saud University - Computer and Information Sciences 34 (2022) 6048–6056. URL: https://www.sciencedirect.com/science/article/pii/S1319157821001804. doi:https://doi.org/10.1016/j.jksuci.2021.07.013.

[19] K. B. Ozler, K. Kenski, S. Rains, Y. Shmargad, K. Coe, S. Bethard, Fine-tuning for multi-domain and multi-label uncivil language detection, in: Proceedings of the Fourth Workshop on Online Abuse and Harms, Association for Computational Linguistics, Online, 2020, pp. 28–33. URL: https://aclanthology.org/2020.alw-1.4. doi:10.18653/v1/2020.alw-1.4.

[20] P. G. Hoang, C. D. Luu, K. Q. Tran, K. V. Nguyen, N. L.-T. Nguyen, ViHOS: Hate speech spans detection for Vietnamese, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 652–669. URL: https://aclanthology.org/2023.eacl-main.47.

[21] X. Ma, End-to-end sequence labeling via bi-directional lstm-cnns-crf, 2016. URL: https://arxiv.org/abs/1603.01354.

[22] M. E. Peters, W. Ammar, C. Bhagavatula, R. Power, Semi-supervised sequence tagging with bidirectional language models, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1756–1765. URL: https://aclanthology.org/P17-1161. doi:10.18653/v1/P17-1161.