LINUS@EEUCA 2026: Fine-grained Toxicity Detection in Gaming Chat using Multilingual Transformers

Prajwal Ghimire
Kathmandu University
prajwalghimire22@gmail.com

Aashish Mahato
Kathmandu University
aashishmtho24@gmail.com

Sunil Regmi
Kathmandu University
sunil.regmi@ku.edu.np

Abstract

The detection of toxic behavior in online gaming communities is crucial for maintaining safe digital spaces, yet remains challenging due to subtle context-dependent and intent-driven language. The GameTox dataset consists of around 53K World of Tanks chat utterances annotated across six categories: Non-toxic, Insults and Flaming, Other Offensive Texts, Hate and Harassment, Threats, and Extremism (Naseem et al., 2025). Our best performing approach, across multiple transformer-based architecture experimentations, is based on the multilingual BERT variant mmBERT-base fine-tuned with class-weighted cross-entropy loss. The best mmBERT-base model achieved a Macro F1 of 0.5882 during validation and an official test Macro F1 of 0.5104 on the shared task leaderboard. An internal held-out evaluation on a development split yielded 0.4282, which we analyze to understand distributional sensitivity to gaming slang and class imbalance. The code is available at: https://github.com/sunilRegmi-ai/eeuca-toxicity-detection.

1 Introduction

The global gaming industry has reached an unprecedented scale. Prior research has explored contextual bandit algorithms to balance exploration and exploitation, optimizing costly real-time toxicity monitoring resources in multiplayer environments where no pre-existing predictive models are available (Morrier et al., 2025). Although chat-enabled titles foster social engagement, they also facilitate toxic behavior and cyberbullying, particularly in team-based competitive environments (Kwak et al., 2015; Naseem et al., 2025).

The detection of hate speech remains a complex societal challenge, with various machine learning and natural language processing architectures being developed to classify diverse forms of online toxicity ranging from religious hate to sexism (Parihar et al., 2021). Early rule-based approaches relied on lexicon and pattern-based annotation systems, which covered only about 16% of distinct vocabulary but over 60% of actual word usage, reflecting the repetitive nature of in-game chat(Märtens et al., 2015). However, such approaches remain limited handling the adversarial nature of in-game slang, where players frequently use creative obfuscation, special character substitutions, and evolving abbreviations to circumvent moderation, prompting the development of robust character-level CNN and hybrid transformer models (Lee et al., 2025). Content moderation systems of this kind require responsible usage, as LLMs can inherit and amplify biases present in training data, necessitating community engagement and bias mitigation to ensure fairness in public discourse and policy-making (Thapa et al., 2025b). Thus, modern dataset creation is actively recognizing the critical role of the author's behavioral intent rather than just surface level lexical features to align automated moderation with actual community policies (Wang et al., 2024). These methodologies are highly relevant to the goals of ToxIntent@EEUCA 2026, which aims to understand toxic behavior in gaming communities to promote healthier digital spaces (Thapa et al., 2026; Hürriyetoğlu et al., 2026).

The GameTox dataset (Naseem et al., 2025) provides around 53K chat utterances from *World of Tanks* annotated across six fine-grained categories: Non-toxic (0), Insults and Flaming (1), Other Offensive Texts (2), Hate and Harassment (3), Threats (4), and Extremism (5). The annotation schema used for this task shares its foundational principles with the multimodal CrisisHateMM dataset (Bhandari et al., 2023), which highlights the importance of distinguishing between directed and undirected toxic intent. Our approach is consistent with findings regarding the reliability of models such as FastText and BERT while BERT's contextual embeddings are better suited to identifying and distinguishing complex targets, such as indi-

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 216–222

viduals, organizations, and communities whereas FastText was found to be exceptional for language identification and was less reliable for the more nuanced task of target identification (Acharya et al., 2025). Prior work on fine-grained target classification in hate speech detection demonstrates that contextual embeddings support models to distinguish between nuanced category types, including individual, community, and organizational targets across Devanagari-script languages (Thapa et al., 2025a).

The monolingual pre-training on massive domain-specific corpora has shown superior results in capturing rich semantics and grammatical structures for specific low-resource scripts like Nepali (Timilsina et al., 2022; Maskey et al., 2022), yet we opt to leverage multilingual mmBERT variants to ensure robust feature representation across diverse gaming communities, building on the proven workings of domain adapted transformers retrained on abusive corpora (Caselli et al., 2021).

However, mmBERT variants remain vulnerable to the linguistic volatility of gaming slang. Furthermore, reliance on machine-translated non-English samples introduces semantic misalignment and translation noise, which can degrade classification performance, particularly in informal contexts (Lamin and Aziz, 2025). Therefore, existing approaches struggle to generalize across rare toxicity categories under severe class imbalance and distributional shift, where intent is subtle and context-dependent.

Our participation in ToxIntent@EEUCA 2026 secured a rank of 18th out of 35 competing teams. As a shared-task system description, our primary contributions are a systematic benchmark of five multilingual transformer encoders on a severely class-imbalanced gaming-domain dataset, and an error analysis demonstrating that class-weighted loss alone is insufficient for extremely rare toxicity categories. Our findings reveal that mmBERT-base, optimized with balanced class-weighted cross-entropy loss, substantially outperforms other multilingual encoder baselines on validation and achieves an official test Macro F1 of 0.5104. An internal dev-split evaluation of 0.4282 further reveals sensitivity to distributional shift in rare toxicity categories such as *Extremism* and *Threats*, where players actively employ irony, sarcasm, and localized slang to disguise toxic intent. The code is available at :`https://github.com/sunilRegmi-ai/eeuca-toxicity-detection`.

2 Background

Modern Natural Language Understanding (NLU) is grounded in deep bidirectional pre-training, allowing models to jointly condition on both left and right context in all layers (Devlin et al., 2019). Slot-gated architectures have proven reliable for mapping specific relationships between intent categories and semantic slots by utilizing a slot gate that optimizes the global relationship between intent and slot attention vectors (Goo et al., 2018). This paradigm has been successfully extended via joint fine-tuning strategies to unify classification and slot filling using a shared representation (Chen et al., 2019). Domain-adapted models like HateBERT utilize retraining on curated datasets of banned communities to improve performance on out-of-distribution toxic text to improve the detection of rare and highly offensive categories (Caselli et al., 2021). Frameworks such as ToXCL unify detection with explanation generation using knowledge distillation from a teacher classifier to mitigate error propagation and provide transparency in the automated moderation of implicit hate (Hoang et al., 2024).

The impact of Large Language Models (LLMs) has marked a new era in Computational Social Science (CSS), offering the capacity to interpret human communication nuances and patterns in ideological shifts (Thapa et al., 2025b). However, substantial generalization gaps often remain, as existing top-tier models continue to struggle with the added complexity of diverse NLU benchmarks spanning inference, similarity, and masked evaluation (Nyachhyon et al., 2025). Research on low-resource scripts reveals that dedicated, fine-tuned masked language models (like NepBERTa) frequently outperform generalized LLMs on sequence tagging tasks such as Named-Entity Recognition (NER) and POS tagging (Subedi et al., 2024).

A strategic research on informal digital discourse further shows that multi-aspect annotation schemes uncover nuanced layers of intent such as profanity, violence, feedback and sarcasm overlooked by basic binary systems (Singh et al., 2020). This is particularly evident in anti-establishment discourse and election-related text, where multi-target classification (e.g., individual vs. community) is crucial for understanding the propagation of hate (Rauniyar et al., 2023; Thapa et al., 2023). Bidirectional LSTM modeling paired with appropriate word embeddings remains highly effective for separating

offensiveness and profanity into distinct classification tasks, even amid the noise of informal social media text for specific behaviors like profanity detection (Adhikari et al., 2024).

The necessity of specialized encoder strategies to manage domain-specific linguistic complexity is underscored in applications such as legal machine translation, where custom-built parallel corpora are required to ground encoder-decoder architectures against data sparsity (Poudel et al., 2024). This principle informs our architectural adaptation for the slang-heavy gaming domain. Cross-lingual challenges including semantic misalignment and translation noise remain pertinent when modeling the multilingual roots of global subcultures (Lamin and Aziz, 2025). Finally, empirical evaluations confirm that fine-tuned lightweight transformers (like DistilBERT) continue to provide optimal accuracy-cost trade-offs in continuous gaming moderation when compared to the computational expense of large generative LLMs or Retrieval-Augmented Generation (RAG) pipelines (Tereshchenko and Hämäläinen, 2025).

3 Dataset and Task

The Shared Task on Understanding Toxic Behavioral Intent in Gaming Chat Logs utilizes the GameTox dataset (Naseem et al., 2025). The dataset comprises around 53K chat utterances sourced from the multiplayer online game *World of Tanks*. The primary objective of this task is to capture the complex relationship between user intent and linguistic features across a fine-grained classification task across six categories.

Class ID	Category	Instances
0	Non-toxic	43,497
1	Insults and Flaming	7,407
2	Other Offensive Texts	2,343
3	Hate and Harassment	349
4	Threats	75
5	Extremism	30

Table 1: Distribution of the dataset

The vast majority of utterances fall into the Non-toxic class, while severe toxicity categories, such as Threats and Extremism, are extremely rare that shows severe class imbalance. The gaming specific slang, obfuscation techniques, and informal syntax, also creates a highly noisy and adversarial text environment. The official evaluation metric for the shared task is the Macro F1-score, which weighs the performance across all classes equally, heavily penalizing models that overfit to the majority class.

4 Methodology

Our system approaches the toxicity classification task with multiple experimental tracks explained below.

4.1 Model Architectures and Training Setup

We systematically benchmarked several modeling paradigms. In the initial phase, we evaluated a broad set of multilingual encoder models — ToxicXLM-RoBERTa, XLM-RoBERTa, m-DistilBERT, and m-BERT — using the Hugging Face Trainer API (Wolf et al., 2020). The raw chat utterances are loaded from CSV files and merged on a common index column, with the text field standardized and labels cast to integers across all splits. Each input is tokenized using the model's native tokenizer with truncation and padding to a maximum sequence length, producing fixed-length token sequences fed directly into the classification head. Class weights are computed from the training label distribution using scikit-learn's `compute_class_weight` and passed to a customized `WeightedTrainer` subclass that overrides the default cross-entropy loss to address class imbalance. All non-DeBERTa models are trained with FP16 mixed precision when CUDA is available. The implementation relies on the Hugging Face `transformers` and `datasets` libraries, with `pytorch` as the framework, `scikit-learn` for class weight computation, and `accelerate` for distributed training support. The full implementation details and reproducibility instructions are available at: `https://github.com/sunilRegmi-ai/eeuca-toxicity-detection`.

The experimental logs in the subsequent phase identified mmBERT variants as the superior architecture due to its pre-training on massively multilingual social media and informal web corpora (Marone et al., 2025). The mmBERT-base variant was fine-tuned using the Hugging Face Trainer API with a customized WeightedTrainer applies class-weighted cross-entropy loss to address class imbalance. An initial run with a learning rate of 3e-06, batch size of 32, and sequence length of 64 yielded a validation Macro F1 of 0.4933. A subsequent run with an increased learning rate of 1e-05, batch size of 64, and reduced sequence length of 32 produced a substantial improvement and achieved a valida-

tion Macro F1 of 0.5882 — a gain of ∼0.09 that underscores the sensitivity of mmBERT to learning rate and batch size scaling. Both runs used weight decay of 0.01 and early stopping with a patience of 3 epochs, with the best checkpoint selected based on validation Macro F1.

5 Results and Discussion

The evaluation metrics for our primary transformer benchmarking experiments are listed in Table 2.

Model	Val F1	Test F1	Test Acc
Toxic-XLM-RoBERTa	0.3558	0.3520	0.8281
XLM-RoBERTa	0.3830	0.3839	0.8130
m-DistilBERT	0.3907	0.3578	0.7942
m-BERT	0.4146	0.4243	0.8249
mmBERT-base	**0.5882**	**0.5104**	**0.8716**

Table 2: Experimental Results on Validation and Test Sets

These results, reported on the official validation and test splits, identify mmBERT as the most reliable base architecture for this specific dataset. The standard generalized models and domain-specific toxicity variants like Toxic-XLM-RoBERTa struggled to surpass the ∼0.42 Macro F1 barrier. Our best mmBERT-base model achieved a validation Macro F1 of 0.5882 and an official leaderboard Test Macro F1 of **0.5104** securing a rank of 18th out of 35 participating teams. An additional internal evaluation on a held-out portion of the development set yielded a Test Macro F1 of 0.4282; this lower figure reflects sensitivity to the specific distributional characteristics of that split rather than the true held-out test performance. The ablation between the two mmBERT-base runs further confirms that scaling the learning rate from 3e-06 to 1e-05 and the batch size from 32 to 64 were the primary determinants of the ∼0.09 validation F1 improvement. The two runs also differ in sequence length (64 vs. 32), so this hyperparameter was co-varied and cannot be isolated from the ablation alone. However, both the token length distribution of the training split and the nature of the dataset provide strong empirical justification: the 75th, 90th, and 95th percentiles of whitespace-tokenized utterance lengths are 4, 6, and 8 tokens respectively, with a mean of 3.02 and a maximum of 30 tokens, confirming that a `max_length` of 32 provides complete coverage for all training utterances. This is consistent with the source domain: Naseem et al. (2025) collected utterances from World of Tanks real-time in-game chat, where annotation guideline examples reach a maximum of 6 whitespace tokens, and the most discriminative toxic and game-slang tokens identified in their dataset are predominantly single-word items (e.g., *die*, *cancer*, *kill*), confirming that toxicity signal in this domain is lexically concentrated rather than requiring long contextual spans.

5.1 Error Analysis

Our official test Macro F1 of 0.5104 reflects a moderate ∼0.08 gap from the validation score of 0.5882. A wider gap of ∼0.16 was observed on an internal development split evaluation. This generalization gap underscores the core difficulty of this shared task: extreme out-of-distribution linguistic volatility.

Error analysis indicates that the minority classes (*Extremism* and *Threats*), comprising only 30 and 75 instances respectively out of 53,701 total, suffer heavily from false negatives. The extreme scarcity of these categories means the model has insufficient exposure to their linguistic patterns during training despite balanced class weighting. Players frequently disguise severe intent using irony, sarcasm, inside jokes, and highly localized slang that models struggle to interpret without broader context. The distributional shift between the validation and unseen test interactions further compounds this, as gaming communities continuously evolve new obfuscation strategies that fall outside the training distribution. Future iterations should explore few-shot augmentation strategies for rare categories and more context-aware architectural solutions to address active linguistic obfuscation.

6 Conclusion

This paper presents our system submission to ToxIntent@EEUCA 2026, a shared task on fine-grained toxicity detection in gaming chat logs using the GameTox dataset (Naseem et al., 2025). As a shared-task system description, our primary contributions are a systematic benchmark of five multilingual transformer encoders on a severely class-imbalanced gaming-domain dataset, and an error analysis demonstrating that class-weighted loss alone is insufficient for extremely rare toxicity categories. We identified mmBERT-base (Marone et al., 2025) as the most effective architecture for this task, owing to its pre-training on massively multilingual social media and informal web corpora. Our

best mmBERT-base model fine-tuned with class-weighted cross-entropy loss and optimized hyperparameters resulted in validation Macro F1 of 0.5882 and an official test Macro F1 of 0.5104. Our ablation analysis further demonstrates that learning rate and batch size scaling are critical determinants of performance gain for mmBERT on this task. The substantial ~ 0.08 F1 generalization gap between validation and test environments underscores the inherent difficulty of detecting fine-grained toxic intent in adversarial, slang-heavy gaming discourse. Future work should explore few-shot augmentation for rare toxicity categories, context-aware architectures, and ensemble strategies to bridge this generalization gap.

Limitations

The severe class imbalance in the GameTox dataset (Naseem et al., 2025), particularly for *Extremism* (30 instances) and *Threats* (75 instances), limits the model's exposure to rare toxicity patterns, and balanced class weighting alone is insufficient to fully compensate for this scarcity. Our evaluation is limited to the GameTox dataset sourced from a single game (*World of Tanks*), and generalization to other gaming communities or platforms remains unverified. The absence of explicit data augmentation or few-shot strategies for minority classes is a known weakness of our current pipeline. Finally, the distributional shift between validation and test interactions suggests that our models are sensitive to evolving gaming slang and obfuscation strategies that fall outside the training distribution, a challenge that requires more robust cross-domain generalization techniques in future iterations.

Acknowledgments

We thank the organizers of ToxIntent@EEUCA 2026 (Thapa et al., 2026; Hürriyetoğlu et al., 2026) for providing the datasets and their support throughout this research.

References

Darwin Acharya, Sundeep Dawadi, Shivram Saud, and Sunil Regmi. 2025. Paramananda@NLU of Devanagari script languages 2025: Detection of language, hate speech and targets using FastText and BERT. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 334–338.

Abiral Adhikari, Prashant Manandhar, Reewaj Khanal, Samir Wagle, Praveen Acharya, and Bal Krishna Bal. 2024. Profanity and offensiveness detection in Nepali language using bi-directional LSTM models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 515–521.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Tommaso Caselli, Valerio Basile, Jelena Mitrovi'c, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25.

Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *Preprint*, arXiv:1902.10909.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Chih-Wen Goo, Guang Gao, Yun-Kai Hsu, and 1 others. 2018. Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the NAACL: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Nhat M. Hoang, Xuan Long Do, Duc Anh Do, Duc Anh Vu, and Luu Anh Tuan. 2024. ToXCL: A unified framework for toxic speech detection and explanation. In *Proceedings of the 2024 Conference of the ACL: Human Language Technologies (Volume 1: Long Papers)*, pages 6460–6472.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, Laxmi Thapa, and Surabhi Adhikari. 2026. Overview of the workshop on event extraction and understanding: Challenges and applications. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*, page 3739–3748. Association for Computing Machinery.

Nor Zakiah Lamin and Azwa Abdul Aziz. 2025. Cross-lingual sentiment analysis in low-resource languages: A recent review on tasks, methods and challenges. *International Journal of Advanced Computer Science and Applications*.

Jaehong Lee, Pavinee Rerkjirattikal, and Sanggyu Nam. 2025. Toxic chat detection in online games using

hybrid bert and character-level cnn. In *MakeLearn, TIIM & PIConf 2025: Accelerated Innovation (AI); Sustainability for Better Humanity*. ToKnowPress.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel, and Bal Krishna Bal. 2022. Nepali encoder transformers: An analysis of auto encoding transformer language models for Nepali text classification. In *Proceedings of the 1st Annual Meeting of the SIGUL*, pages 106–111.

Jacob Morrier, Rafal Kocielnik, and R. Michael Alvarez. 2025. Bandit algorithms for efficient toxicity detection in competitive online video games. *IEEE Access*, 13:103109–103117.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6.

Usman Naseem, Shuvam Shiwakoti, Siddhant Bikram Shah, Surendrabikram Thapa, and Qi Zhang. 2025. GameTox: A comprehensive dataset and analysis for enhanced toxicity detection in online gaming communities. In *Proceedings of the 2025 Conference of the NAACL: Human Language Technologies (Volume 2: Short Papers)*, pages 440–447.

Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. 2025. Consolidating and developing benchmarking datasets for the Nepali natural language understanding tasks. In *Proceedings of the 14th IJCNLP and the 4th AACL*, pages 1906–1925.

Anil Singh Parihar, Surendrabikram Thapa, and Sushruti Mishra. 2021. Hate speech detection using natural language processing: Applications and challenges. In *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 1302–1308. IEEE.

Shabdapurush Poudel, Bal Krishna Bal, and Praveen Acharya. 2024. Bidirectional English-Nepali machine translation(MT) system for legal domain. In *Proceedings of the 3rd Annual Meeting of the SIGUL @ LREC-COLING 2024*, pages 53–58.

Kritesh Rauniyar, Sweta Poudel, Shuvam Shiwakoti, Surendrabikram Thapa, and 1 others. 2023. Multi-aspect annotation and analysis of nepali tweets on anti-establishment election discourse. *IEEE Access*, 11:143092–143115.

Oyesh Mann Singh, Sandesh Timilsina, Bal Krishna Bal, and Anupam Joshi. 2020. Aspect based abusive sentiment detection in nepali social media texts. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 301–308.

Bipesh Subedi, Sunil Regmi, Bal Krishna Bal, and Praveen Acharya. 2024. Exploring the potential of large language models (LLMs) for low-resource languages: A study on NER and POS tagging for Nepali language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics (LREC-COLING 2024)*, pages 6974–6979.

Yehor Tereshchenko and Mika K Hämäläinen. 2025. Efficient toxicity detection in gaming chats: A comparative study of embeddings, fine-tuned transformers and llms. *Journal of Data Mining & Digital Humanities*.

Surendrabikram Thapa, Rauniyar Kritesh, Shiwakoti Shuvam, and 1 others. 2023. Nehate: Large-scale annotated data shedding light on hate speech in nepali local election discourse. In *ECAI 2023*, pages 2346–2353. IOS Press.

Surendrabikram Thapa, Kritesh Rauniyar, Farhan Ahmad Jafri, Surabhi Adhikari, Kengatharaiyer Sarveswaran, Bal Krishna Bal, Hariram Veeramani, and Usman Naseem. 2025a. Natural language understanding of Devanagari script languages: Language identification, hate speech and its target detection. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 71–82.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025b. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026. Understanding toxic behavior in gaming communities using ai to promote healthier digital spaces. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Sulav Timilsina, Milan Gautam, and Binod Bhattarai. 2022. NepBERTa: Nepali language model trained in a large corpus. In *Proceedings of the 2nd AACL and 12th IJCNLP (Volume 2: Short Papers)*, pages 273–284.

Xinyu Wang, Sai Koneru, Pranav Narayanan Venkit, and 1 others. 2024. The unappreciated role of intent in algorithmic moderation of social media content. *Preprint*, arXiv:2405.11030.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.