Mast Kalandar at SemEval-2024 Task 8: On the Trail of Textual Origins: RoBERTa-BiLSTM Approach to Detect AI-Generated Text

Jainit Sushil Bafna
IIIT Hyderabad

jainit.bafna@research.iiit.ac.in

Hardik Mittal[*]
IIIT Hyderabad

hardik.mittal@research.iiit.ac.in

Suyash Sethia[*]
IIIT Hyderabad

suyash.sethia@research.iiit.ac.in

Manish Shrivastava
IIIT Hyderabad

m.shrivastava@iiit.ac.in

Radhika Mamidi
IIIT Hyderabad

radhika.mamidi@iiit.ac.in

Abstract

Large Language Models (LLMs) have showcased impressive abilities in generating fluent responses to diverse user queries. However, concerns regarding the potential misuse of such texts in journalism, educational, and academic contexts have surfaced. SemEval 2024 introduces the task of Multigenerator, Multidomain, and Multilingual Black-Box Machine-Generated Text Detection, aiming to develop automated systems for identifying machine-generated text and detecting potential misuse. In this paper, we i) propose a RoBERTa-BiLSTM based classifier designed to classify text into two categories: AI-generated or human ii) conduct a comparative study of our model with baseline approaches to evaluate its effectiveness. This paper contributes to the advancement of automatic text detection systems in addressing the challenges posed by machine-generated text misuse. Our architecture ranked 46th on the official leaderboard with an accuracy of 80.83 among 125.

1 Introduction

The task of classifying text as either AI-generated or human-generated holds significant importance in the field of natural language processing (NLP). It addresses the growing need to distinguish between content created by artificial intelligence models and that generated by human authors, a distinction crucial for various applications such as content moderation, misinformation detection, and safeguarding against AI-generated malicious content. This task is outlined in the task overview paper by (Wang et al., 2024), emphasizing its relevance and scope in the NLP community.

Our system employs a hybrid approach combining deep learning techniques with feature engineering to tackle the classification task effectively. Specifically, we leverage a BiLSTM (Bidirectional Long Short-Term Memory) (Schuster and Paliwal, 1997) neural network in conjunction with RoBERTa (Liu et al., 2019), a pre-trained language representation model, to capture both sequential and contextual information from the input sentences. This hybrid architecture enables our system to effectively capture nuanced linguistic patterns and semantic cues for accurate classification.

Participating in this task provided valuable insights into the capabilities and limitations of our system. Quantitatively, our system achieved competitive results, ranking 46 relative to other teams in terms of accuracy and F1 score. Qualitatively, we observed that our system struggled with distinguishing between sentences generated by AI models trained on specific domains or datasets with highly similar linguistic patterns.

We have released the code for our system on GitHub[1], facilitating transparency and reproducibility in our approach.

2 Related Works

In the field of detecting machine-generated text, numerous methodologies and models have been examined. A distinguished methodology is the application of the RoBERTa Classifier, which enhances the RoBERTa language model through fine-tuning for the specific purpose of identifying machine-generated text. The proficiency of pre-trained classifiers like RoBERTa in this domain has been affirmed through various studies, including those conducted by (Solaiman et al., 2019) and additional research by (Zellers et al., 2019; Ippolito et al., 2019; Bakhtin et al., 2020; Jang et al., 2020; Uchendu et al., 2021). Concurrently, the XLM-R Classifier exploits the multilingual training of the XLM-RoBERTa model to effectively recognize machine-generated text in various languages, as demonstrated by (Conneau et al., 2019).

---

[*]Equal contribution.

[1]https://github.com/Mast-Kalandar/SemEval2024-task8

Model/Source	chatGPT	cohere	davinci	dolly	human
wikihow	3000	3000	3000	3000	15499
wikipedia	2995	2336	3000	2702	14497
reddit	3000	3000	3000	3000	15500
arxiv	3000	3000	2999	3000	15498
peerread	2344	2342	2344	2344	2357

a)

Model/Source	bloomz	human
wikihow	500	500
wikipedia	500	500
reddit	500	500
arxiv	500	500
peerread	500	500

b)

Table 1: Table a) contains statistics about the train split. Table b) contains statistics about the validation split from the M4 dataset

Alternatively, the exploration of logistic regression models that incorporate GLTR (Giant Language model Test Room) features has been undertaken. These models strive to discern subtleties in text generation methodologies by analyzing token probabilities and distribution entropy, as investigated by (Gehrmann et al., 2019). Furthermore, detection efforts have utilized stylometric and NELA (News Landscape) features, which account for a broad spectrum of linguistic and structural characteristics, including syntactic, stylistic, affective, and moral dimensions, as reported by (Li et al., 2014) and (Mitchell et al., 2023). Additionally, proprietary frameworks like GPTZero, devised by Princeton University, focus on indicators such as perplexity and burstiness to analyze texts for machine-generated content identification. Although the specific technical details are sparingly disclosed, the reported effectiveness of GPTZero in identifying outputs from various AI language models highlights its significance in the ongoing development of machine-generated text detection strategies (Ouyang et al., 2022; Brown et al., 2020; Radford et al., 2019; Touvron et al., 2023).

3 Background

3.1 Dataset

For the machine-generated text, the researchers used various multilingual language models like ChatGPT(OpenAI, 2024), textdavinci-003(OpenAI, 2022), LLaMa(Touvron et al., 2023), FlanT5(Chung et al., 2022), Cohere(Cohere, 2024), Dolly-v2(databricks, 2022), and BLOOMz(Muennighoff et al., 2023). These models were given different tasks like writing Wikipedia articles, summarizing abstracts from arXiv, providing peer reviews, answering questions from Reddit and Baike/Web QA, and creating news briefs. As evident from Table 1, the training set lacks any sentences generated by the Bloomz model, which stands as the sole model represented in the validation set. This deliberate choice ensures a robust assessment of our model's generalization capabilities across all machine-generated outputs, regardless of the specific model generating them. By exposing our model to diverse machine-generated sentences during training, including those from unseen models like Bloomz in the validation set, we aim to evaluate its ability to effectively generalize to novel inputs and make reliable predictions across the spectrum of machine-generated text.

3.2 Task

We focused on Subtask-A of the SemEval Task 8 which involves developing a classifier to differentiate between monolingual sentences generated by artificial intelligence (AI) systems and those generated by humans. This classification task is essential for distinguishing the origin of text and understanding whether it was produced by AI models or by human authors.

3.2.1 Objective

The primary objective is to build a robust classifier capable of accurately distinguishing between AI-generated and human-generated sentences. The classifier should generalize well across various AI models and domains, ensuring consistent perfor-

Model	Accuracy	F1	Precision	Recall	Params*
Full RoBERTa fine tune	80.68	80.54	81.55	80.68	124M
LoRA with RoBERTa (Freezed)	81.59	81.06	85.64	**81.59**	0.7M
LoRA with LongFormer	75.34	75.14	76.16	75.34	6M
BiLSTM with RoBERTa (Un-Freezed)	70.77	61.15	**91.19**	46.00	18M
GRU with RoBERTa (Freezed)	74.65	80.54	81.55	80.68	3M
BiLSTM with RoBERTa (Freezed)	**82.52**	**82.14**	83.96	80.40	4M

Table 2: The performance of the models tried on the dev set of the dataset.
*The params only accounts for trainable unfreezed parameters.

mance regardless of the specific model or domain from which the text originates.

The goal was to design a model that not only performs this task with high accuracy but also adapts to various AI models and domains. It's crucial for the classifier to accurately identify the origin of sentences, regardless of the technology used to generate them or their subject matter, ensuring broad applicability and effectiveness

4 System Overview

Based on our observation (See 7), we discovered that language modeling task encodes the various features required for detection of AI written text. So we used pretrained RoBERTa in most of our architectures so exploit this power of language models.

4.1 Full RoBERTa Finetune

The Full RoBERTa(Liu et al., 2019) Finetune model, chosen as our baseline, boasted an extensive architecture and possessed the highest parameter count among the models under evaluation. Serving as a comprehensive starting point, this model allowed us to assess the effectiveness of subsequent enhancements in comparison.

4.2 LoRA with RoBERTa (Frozen)

Incorporating Low Rank Adapters (Hu et al., 2021), we applied fine-tuning techniques to the RoBERTa model while strategically freezing all layers. This approach enabled us to adapt the model to our specific task domain, leveraging pre-trained representations effectively.

4.3 LoRA with LongFormer

The limitation of RoBERTa's context length (max 512 tokens) posed challenges for handling lengthy sentences in our dataset. To address this, we investigated LongFormer (Beltagy et al., 2020), a

model designed to efficiently manage longer contexts. Despite employing LoRA for fine-tuning, the model's performance on the validation set fell short of expectations, indicating potential difficulties in generalization.

4.4 RoBERTa (2 Layers unfreezed) + BiLSTM

Expanding upon RoBERTa's capabilities, we introduced a hybrid architecture by unfreezing two layers and integrating a BiLSTM network (Schuster and Paliwal, 1997). RoBERTa served as the primary encoder for sentence representations, with the subsequent BiLSTM layer trained to classify based on the last hidden state.

4.5 RoBERTa (Frozen) + GRU

In our endeavor to augment RoBERTa's capabilities, we devised a hybrid architecture by integrating a Gated Recurrent Unit (GRU) (Chung et al., 2014) network with the frozen RoBERTa model. Within this framework, RoBERTa served as the encoder for generating sentence representations, while a subsequent GRU layer was incorporated for sequential processing and classification tasks. This amalgamation aimed to leverage the strengths of both RoBERTa's contextual understanding and GRU's recurrent dynamics, contributing to enhanced performance on our target task.

4.6 RoBERTa (Frozen) + BiLSTM

In our pursuit of enhancing RoBERTa's capabilities, we devised a hybrid architecture by coupling a Bidirectional Long Short-Term Memory (BiLSTM) network with the RoBERTa model (Liu et al., 2019). In this setup, RoBERTa functioned as the encoder for sentence representations, while a subsequent BiLSTM layer was employed for classification, utilizing the last hidden state for decision-making. For a detailed visual representation of the model's architecture, please refer to the accompanying Figure
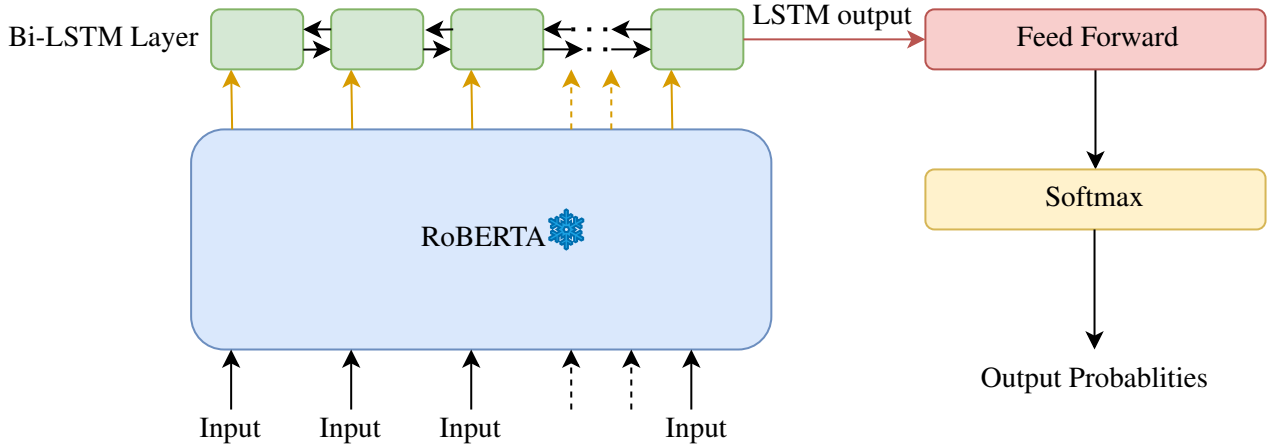
Figure 1: Our proposed architecture of BiLSTM with freezed RoBERTa

1.

We explored various methodologies (refer to Table 2 for detailed performance metrics) before selecting the optimal approach as our final model. Subsequently, we assessed the performance of the chosen model, RoBERTA (Freezed) + BiLSTM, on the test dataset.

5 Experiments

5.1 Preprocessing

All textual data underwent standard preprocessing steps, including tokenization, lowercasing, and punctuation marks. Additionally, specific domain-related preprocessing, such as handling special characters or domain-specific terms, was performed as necessary.

5.2 Hyperparameter Tuning

Hyperparameters were tuned using a combination of grid search and random search techniques. We explored various hyperparameter combinations to identify the optimal configuration for each model variant.

The configuration for LSTM and GRU used in Table 2 is `hidden_size=256`, `layers=2`, `dropout=0.2`, with `LoRA rank being 20` has been found as the best configuration for the models. For RoBERTa+LSTM model's feedforward had a single weight matrix of dimension 512*2.

6 Results

We tested our models on various models on the test set. The results can be viewed in (Table: 3).
Ranking: Our BiLSTM+RoBERTa model achieved a ranking of 46 out of 125 participants

in the competition, demonstrating its competitive performance (as shown in Table 3). These results highlight the effectiveness of various models, including BiLSTM+RoBERTa and GRU+RoBERTa, in addressing the task objectives. We submitted BiLSTM+RoBERTa based on its strong performance on the validation set. However, after testing all models listed in Table 3, we found that GRU+RoBERTa achieved a significantly better result, with an accuracy increase of approximately 4%.

7 Conclusion

In conclusion, our BiLSTM+RoBERTa model effectively tackled the task, achieving competitive results, thanks to its deep learning and pre-trained language model. While a similar model with unfrozen RoBERTa boasted higher precision, its complexity came at the cost of increased parameters.

Impressively, our model ranked 46th out of 125 competition entries (Table 3), showcasing its potential alongside approaches like GRU+RoBERTa. Interestingly, post-competition analysis revealed GRU+RoBERTa's superior accuracy (by about 4%). This highlights the value of exploring diverse architectures and hyperparameter tuning for peak performance.

Moving forward, there are several avenues for future work to explore. Firstly, further experimentation with different model architectures, including alternative combinations of encoders and classifiers, could potentially yield improvements in performance. Additionally, fine-tuning hyperparameters and exploring advanced techniques for model optimization may enhance the robustness and generalization capabilities of our system. Furthermore,

Model	Accuracy	F1	Precision	Recall	Params*
Full RoBERTa fine tune[+]	**88.47**	**88.44**	**93.36**	84.02	124M
LoRA with RoBERTa (Freezed)	80.91	80.18	83.88	80.14	0.7M
LoRA with LongFormer	63.39	57.51	72.45	61.67	6M
BiLSTM with RoBERTa (Un-Freezed)	80.80	80.19	83.08	80.12	18M
GRU with RoBERTa (Freezed)	84.71	84.33	86.53	84.13	3M
BiLSTM with RoBERTa (Freezed)	80.83	80.83	74.65	**96.16**	4M

Table 3: The performance of the models tried on the test set of the dataset.
* The params only accounts for trainable unfreezed parameters.
+ Baseline mentioned in task overview paper

incorporating additional contextual information or domain-specific knowledge could potentially augment the model's understanding and performance on specific tasks. Overall, our findings contribute to the ongoing research efforts in natural language processing and provide valuable insights for future developments in this domain.

References

Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. 2020. Energy-based models for text. *CoRR*, abs/2004.10188.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

Cohere. 2024. Cohere: Chat. https://cohere.com/. Accessed: February 20, 2024.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

databricks. 2022. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM. dolly-v2. Accessed: February 20, 2024.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2019. Human and automatic detection of generated text. *CoRR*, abs/1911.00650.

Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, Sang-ug Kang, and Jong Wook Kim. 2020. Bilstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. *Applied Sciences*, 10(17).

Jenny S. Li, John V. Monaco, Li-Chiou Chen, and Charles C. Tappert. 2014. Authorship authentication using short messages from social networking sites. In *2014 IEEE 11th International Conference on e-Business Engineering*, pages 314–319.

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.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

OpenAI. 2022. text-davinci-003: A Variant of the GPT-3 Language Model. https://openai.com. Accessed: February 20, 2024.

OpenAI. 2024. ChatGPT: A Large-Scale Transformer-Based Language Model. https://openai.com/research/chatgpt. Accessed: February 20, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Mike Schuster and Kuldip Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45:2673 – 2681.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. 2019. Release strategies and the social impacts of language models. *CoRR*, abs/1908.09203.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for turing test in the age of neural text generation. *CoRR*, abs/2109.13296.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection.

In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Malta.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Appendix A

A. Setup

In this study, we implemented a methodology aimed at distinguishing human-generated sentences from machine-generated ones within a training dataset. To achieve this, we initially segregated the dataset into two distinct subsets: one containing human-generated sentences and the other comprising machine-generated ones. Subsequently, we trained separate models utilizing these segregated datasets. Specifically, we employed two distinct models for this task : i) Bidirectional Long Short-Term Memory (**BiLSTM**) model, ii) **RoBERTa** model.

Following the training phase, we proceeded to evaluate the performance of both models on a validation dataset. During this evaluation, we measured the loss incurred by each model when tasked with discerning between human-generated and machine-generated sentences. This evaluation process was crucial for assessing the efficacy and generalization capabilities of the trained models in accurately distinguishing between the two types of sentences.

B. Results

The results are in form of graphs in Figure 2

(a) Model Trained on : Human Sentences, Losses Computed on : Human Sentences

(b) Model Trained on : Human Sentences, Losses Computed on : Machine Sentences

(c) Model Trained on : Machine Sentences, Losses Computed on : Human Sentences

(d) Model Trained on : Machine Sentences, Losses Computed on : Machine Sentences
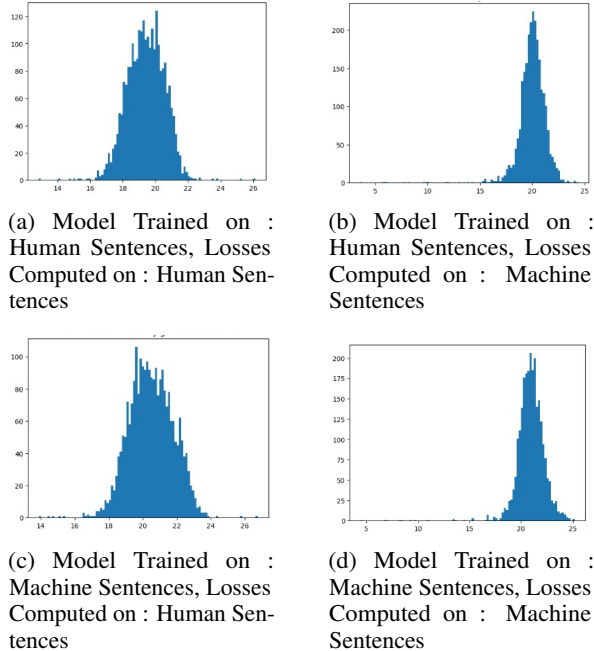
Figure 2: Overall Results on Models trained on Human and Machine Generated Sentences and Losses Calculated on Human and Machine Generated Sentences

We noted a consistent pattern across both sets of models – those trained on human-generated sentences and those trained on machine-generated sentences. Specifically, we observed that the losses incurred by human-generated sentences on the validation set exhibited a wider distribution with higher variance, while the losses associated with machine-generated sentences displayed a narrower distribution with lesser variance.

This observation leads to a compelling inference regarding the predictive nature of the model losses for each type of data. The wider distribution and higher variance in losses for human-generated sentences suggest a greater level of unpredictability associated with these sentences. In contrast, the narrower distribution and lesser variance in losses for machine-generated sentences indicate a higher level of predictiveness in the model's performance on these sentences.

This finding sheds light on the inherent characteristics of human-generated versus machine-generated sentences, particularly regarding their predictability when processed by the trained models. Such insights are crucial for understanding the intricacies of model behavior and the challenges posed by different types of data in natural language processing tasks.