Paraphrasing-Based Data Augmentation for Responsible Personnel Classification in Crane Accidents

Deokyeong Kim[1], Leila Kosseim[2], Hongjo Kim[3], and Jong Won Ma[1]

[1]Dept. of Building, Civil and Environmental Engineering, Concordia University, Montreal, QC, Canada H3G 1M8
[2]Dept. of Computer Science and Software Engineering, Concordia University, Montreal, QC, Canada H3G 2W1
[3]School of Civil and Environmental Engineering, Yonsei University, Seoul, South Korea 03722
deokyeong.kim@mail.concordia.ca, leila.kosseim@concordia.ca, hongjo@yonsei.ac.kr, jongwon.ma@concordia.ca

Abstract –

Crane accidents often result in severe injuries and fatalities, necessitating accurate analysis and responsibility prediction to prevent future incidents. However, the scarcity of crane accident-specific datasets and class imbalance pose challenges to developing robust predictive models. This study addresses these challenges by leveraging paraphrase-based data augmentation methodology to expand a dataset of 480 natural language accident descriptions into a balanced dataset of 900 samples. Using Google's Pegasus paraphrasing model, the augmented data improved model performance, as demonstrated by the F1 score increase from 0.29 to 0.5583. Three pre-trained transformer BERT-based models were fine-tuned on the augmented dataset to evaluate their effectiveness in predicting responsible personnel. The results highlight the effectiveness of the paraphrasing technique and transformer-based models in addressing class imbalance and improving classification accuracy. This research demonstrates the potential of natural language text data in enhancing safety analysis and proposes future directions, such as exploring alternative data augmentation methods, leveraging large language models, and experimenting with domain-specific fine-tuning.

Keywords –

Crane Safety, Natural Language Processing, Deep Learning, Data Augmentation.

1 Introduction

Crane accidents frequently occur on construction sites and, due to their heavy nature, often result in severe injuries and fatalities. Previous studies indicate that these accidents are primarily caused by human errors and miscommunication [1-3]. Forensic analysis after crane accidents plays a crucial role in identifying the sequence of events, root cause and pinpointing the responsible personnel. Identifying responsible personnel in crane accidents is critical for improving workplace safety and accurate liability assessment [4]. This research aims to support such processes by providing structured methods to analyze accident reports, making it easier to derive actionable insights for future safety improvements. This research output can potentially be used as an auxiliary tool to provide additional layer of verification and increase overall efficiency in experts' decision-making process. The construction industry generates significant amounts of text-based data, such as incident reports, safety manuals, and communication logs. Text-based analysis has emerged as a powerful tool in safety domains, including construction and manufacturing, for identifying accident processes [5]. Unlike accident statistics that only provide predefined categories and post-incident outcomes [6], natural language-based accident reports capture richer information about conditions, causes, and injuries [7]. For instance, Ma and Chen analyzed 159 text-based reports to identify accident factors [8], while Kumi et al. classified construction accident types using Korean reports [9]. These studies demonstrate the potential of text-based analysis in uncovering critical insights from unstructured data.

However, analyzing crane accident reports poses significant challenges. First, crane accident-specific datasets are scarce, limiting the ability to develop robust models. And this scarcity in labeled data can introduce overfitting during the model training [10]. Second, class imbalance in available datasets—where roles like "Crane Operator" are overrepresented while others like "Owner/User" are underrepresented—hampers the predictive accuracy of machine learning models.

To address these challenges, this study investigates paraphrase-based data augmentation strategies [11-12] for identifying responsible personnel in crane accidents. Using 480 natural language accident descriptions, the dataset was augmented with Google's Pegasus model [13], resulting in a balanced dataset of 900 samples. This augmentation ensures diverse yet semantically consistent training data, mitigating the impact of class imbalance.

Additionally, three pre-trained BERT-based models [14] were fine-tuned on the dataset to evaluate their performance in predicting responsible personnel. This approach aligns with findings by Devlin et al. (2019), which highlight that fine-tuning the entire model on task-specific data, even with limited examples, can significantly enhance performance compared to using fixed embeddings [14].

2 Data

2.1 Data Preprocessing

The original dataset contains 710 crane accident records provided by CRL (Crane Risk Logic, Inc.), where each record includes surveys of the accident descriptions, primary responsible personnel, and more. After removing empty contents and where the target variable (responsible personnel) is not available, a total of 480 crane accident descriptions and their primary responsible personnel were extracted from the original dataset. In this study, we removed or substituted any sensitive information (including personnel names, specific locations, and company identifiers) to protect the privacy of all parties involved. Each crane accident description contains an average of 93 words, ranging from two to three lines to 300 words. There are nine classes for the primary responsible personnel, with 'Crane Operator' being the most frequent class (100 records), while the 'Owner/User' class has the fewest (14 records) (Table 1).

Table 1. Original data per class distribution (total 480)

Class	Name	# of Data
0	Crane Operator	100
1	Rigger	96
2	Lit Director	94
3	Site Supervisor	58
4	Manufacturer	39
5	Other	27
6	Mechanical/Maintenance Issue	27
7	Signal Person	25
8	Owner/User	14

2.2 Input data Preparation

2.2.1 Paraphrase

Due to the imbalanced nature of the original data, this study adopted a paraphrasing technique for input dataset generation. The overall process of data augmentation is shown in Figure 1. A total of five datasets were used in this study. The first dataset contains only the original data, comprising 480 crane accident descriptions. The second dataset is augmented using Google's PEGASUS (Pre-training with Extracted Gap-sentences for Abstractive Summarization) paraphrasing model [13] . This ensures that each class contains 100 balanced samples (Figure 2), resulting in a total of 900 samples for Dataset 2. The third dataset is also generated using paraphrasing; however, the test set was removed in advance before paraphrasing to prevent any potential data leakage during training. For fair comparison, the test dataset size was fixed at 63 samples and remained the same across all three datasets. For the imbalanced original dataset, where the lowest class has only 14 samples, 50% of the smallest class's data count was used as the criterion for constructing the test set. As a result, 7 samples per class were selected, totalling 63 samples for the test set. Table 1 shows the original data distribution per class, and Table 2 data splits for each dataset.

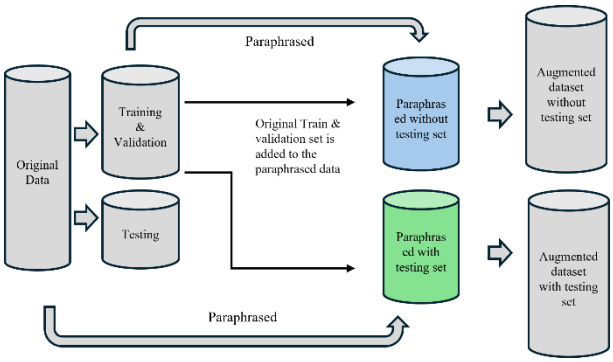

Figure 1. Data split process

Table 2. Description of five dataset used and dataset size

Dataset	Is Test set used for paraphrasing?	Is Test set balanced?	Data size (train-validation, test)
1	No	Yes	(417, 63)
2	No	Yes	(837,63)
3	Yes	Yes	(837,63)
4	No	No	(417, 63)
5	Yes	No	(837,63)

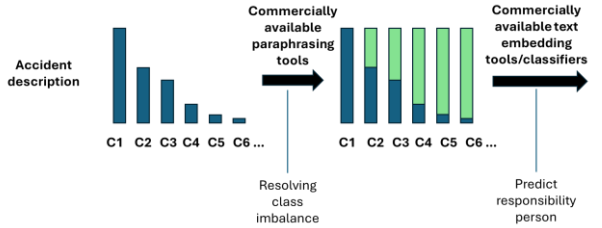

Figure 2. Data augmentation overview

2.2.2 Hyperparameters

The length of the paraphrased text was set to a

minimum of 80% and a maximum of 120% of the original target text. Top-p sampling was used as the sampling method, applied with a cumulative probability threshold (top_p=0.9) to ensure that words were sampled from the most probable token distribution while maintaining linguistic coherence [15]. Only the top 90% of cumulative probability words were considered, and the next word was selected based on normalized probabilities among these options.

In our paraphrase generation task, the temperature parameter was set to 1.5 to control the randomness of token selection during sampling. A temperature value of 1 results in a neutral probability distribution for next-word selection, while values greater than 1 increase randomness by making the probability distribution more uniform. This adjustment enhances output diversity but may occasionally reduce coherence or grammatical accuracy [16].

2.2.3 Paraphrasing quality evaluation

In this study, BLEU (Bilingual Evaluation Understudy) [17] and BERTScore [18] were used to evaluate the quality of paraphrasing. BLEU is a widely used evaluation metric in Natural Language Processing (NLP), which calculates a score based on n-gram overlaps through simple mathematical computations [17]. However, it fails to account for semantic meaning and diversity in paraphrased texts [19]. In contrast, BERTScore leverages embedding vectors to compare the contextual meaning of words, making it well-suited for evaluating semantic similarity [18]. Table 3 provides two paraphrase examples for the 'Owner/User' class. The second paraphrased example in Table 3 scored 0.1163 for BLEU and 0.9162 for BERTScore. For each paraphrased text, both BLEU and BERTScores were calculated. These examples demonstrate the trade-off between semantic similarity and diversity, as reflected in the scores. The resulting paraphrased sentences are compared to its original reference data. The BLEU scores are relatively low due to the diverse rephrasing of the original text, the BERTScores remain high (above 0.8), indicating that the semantic meaning of the text is well-preserved. The average BLEU and BERTScore for 420 paraphrased texts using all the original dataset (excluding 'Crane Operator' class) scored 0.0689 and 0.9181 respectively. This is Calculated to see how diverse the paraphrased result is (lower BLEU) in terms of sentence structure and words choices but still have strong semantic similarity with high BERTScore.

Table 3. Example of two paraphrasing result of 'Owner/User' class with its BLEU, BERTScore

Natural Language Crane Accident Description ('Owner/User' class)	BLEU	BERTScore
Original: A crew was dispatched to an off shore platform to perform prep work for a move to another site. A pedestal crane on site was being used to lift various heavy mechanical equipment onto the deck of the platform. During one lift the load was raised and beginning to swing when the load line broke dropping the load and killing a worker.	-	-
Paraphrased 1: There was work done ahead of a move to the off shore platform. with a team. Heavy mechanical equipment was loaded onto a crane to be put onto adeck the platform. While it was a light Lifting at a construction project one lift had a load line break resulting in a dead worker when it swung.	1.9561E-78	0.8919
Paraphrased 2: A crew was sent to an off shore platform so they could help relocate some structures. It was a very cumbersome lift using a pedestal crane to get heavy mechanical equipment from one location to another. A worker was killed as the loaded load was cut back as it swung by a malfunctioning load line.	0.1163	0.9162

3 Methodology

3.1 Transformers

This study utilized Transformer models, particularly BERT (Bidirectional Encoder Representations from Transformers)-based models, due to its proven effectiveness in NLP tasks [14]. Three types of BERT-based models were used: BERT, RoBERTa (Robustly Optimized BERT Pretraining Approach), and DistilBERT (Distilled Bidirectional Encoder Representations from Transformers). The BERT model, developed by Google, was used as the baseline for

comparison due to its effectiveness in NLP tasks, leveraging bidirectional context to learn semantic meaning [14]. The RoBERTa model, developed by Facebook AI, is an enhanced version of BERT that was trained for a longer duration on larger datasets with improved pretraining techniques [20]. Meanwhile, DistilBERT is a lightweight model that is 40% smaller than BERT but generally achieves competitive performance with reduced computational costs [21].

Each model takes input_ids, which are tokenized text data mapped to integer sequences, along with supplementary information, such as attention_masks, as input. These inputs are passed through the encoder of the transformer model, which generates high-level vector embeddings containing contextual information about the input text using self-attention layers. The self-attention mechanism allows tokens to capture their contextual relationships within the text. For the classification task, a classifier is attached to the final layers of the models [22]. While the encoder component is pretrained, the classifier is initialized randomly. During training, the parameters of both the encoder and classifier are updated to fine-tune the model for the specific task.

3.2 Grid Search

Grid search was conducted to tune hyperparameters, including learning rate, batch size, and the number of epochs. Table 4 summarizes the combinations chosen for each model, for the following hyperparameter ranges explored: learning rate [1e-5, 2e-5, 3e-5], batch size [8, 16], and epochs [5, 10, 15].

3.3 Metrics

Precision, recall, and F1 scores were calculated as evaluation metrics for datasets 1, 2, and 3. F1 score, being the harmonic mean of precision and recall, offers a balanced measure of classification performance. The weighted F1 score is used for imbalanced datasets 4 and 5.

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

$$Weighted\ F1\ Score = \sum_{i=1}^{N} w_i \times F1\ Score_i \quad (4)$$

$$where\ w_i = \frac{number\ of\ samples\ in\ class\ i}{total\ number\ of\ samples}$$

3.4 Input data augmentation: Paraphrasing technique

This study conducted two experiments to validate the effectiveness of paraphrasing in augmenting an imbalanced dataset under varying distribution conditions.

3.4.1 Experiment 1: Balanced distribution

For Experiment 1, Datasets 1,2 and 3 are used. Dataset 1 is the original imbalanced dataset containing total of 480 crane accident descriptions. Dataset 2 is augmented using paraphrasing; however, the 63 test is excluded in paraphrasing. Dataset 3 is augmented with all the data used for paraphrasing. Thus, after data augmentation using paraphrasing for dataset 2 and 3, is balanced with each containing 100 data per class.

3.4.2 Experiment 2: Original skewed distribution

For Experiment 2, Datasets 4 and 5 were used. Dataset 4 contains an imbalanced test set with 63 samples. The test set reflects the original dataset's imbalanced nature by using the stratify option in the train_test_split function to preserve the same class distribution as the original data. The remaining data was used for training and validation purposes. Dataset 5 also contains an imbalanced test set; however, the training and validation data were augmented. This augmentation was performed using the training and validation data from Dataset 4, excluding the test set, to create a larger and more diverse dataset for model training.

4 Result

Table 4 shows the performance of the prediction models on five input datasets for three different BERT-based models. For the original imbalanced data (Dataset 1) with a balanced test set, the best performance was achieved with DistilBERT, scoring 0.2898. With the addition of paraphrasing (Dataset 2), the performance improved across all three models, with RoBERTa showing the most significant improvement, increasing from 0.2544 to 0.3261. The results for Dataset 3 showed further improvement, with DistilBERT achieving the highest score of 0.5583. The confusion matrix in Figure 3 revealed that the model performed well in identifying major classes such as 'Crane Operator (class 0),' with high true positive rates, however failed with underrepresented classes (class 4-8). The confusion matrix for dataset 2 (Figure 4), shows improvement compared to dataset 1(Figure 3).

The precision of the base BERT model showed only a slight increase from 0.2672 to 0.2677 after data augmentation. However, the recall improved from

0.2857 to 0.3333, suggesting that the augmentation enriched the positive examples in this class, enabling the model to generalize better and reduce false positive predictions. In contrast, the DistilBERT model experienced a decline in precision from 0.3714 to 0.3089 when trained on the augmented dataset. This result indicates that lightweight models like DistilBERT may have a lower capacity to generalize effectively when exposed to more diverse data.

However, as indicated in Figure 5, there is evidence of overfitting, as lower-data-count classes were classified almost perfectly in the test set. This is likely because of a certain degree of data leakage, given that the training dataset includes paraphrased samples derived from the entire dataset, thus introducing contexts similar to those in the test data. Further experiments are needed to address these issues, as all models exhibited high performance (F1 scores above 0.6) during the validation process. Future research should explore alternative methodologies to enhance dataset quality by determining the most effective data augmentation strategies and their impact on model performance.

Table 4. Result of 5 datasets with each three BERT based transformer models

Data	Model	Precision	Recall	F1	Best Hyperparameters (learning rate, batch size, epochs)
1	BERT	0.2672	0.2857	0.2159	(2e-5, 8, 10)
	RoBERTa	0.2849	0.3175	0.2544	(3e-5, 16, 10)
	DistilBERT	0.3714	0.3492	0.2898	(2e-5, 8, 15)
2	BERT	0.2677	0.3333	0.2502	(2e-5, 8, 10)
	RoBERTa	0.3687	0.3651	**0.3261**	(3e-5, 16, 10)
	DistilBERT	0.3089	0.3651	0.2979	(3e-5, 8, 10)
3	BERT	0.5897	0.5556	0.5423	(2e-5, 8, 10)
	RoBERTa	0.6713	0.4603	0.4473	(2e-5, 8, 15)
	DistilBERT	0.5597	0.5873	**0.5583**	(3e-5, 8, 15)
Data	Model	Precision	Recall	Weighted F1	Best Hyperparameters (learning rate, batch size, epochs)
4	BERT	0.2810	0.2507	0.2445	(3e-5, 8, 5)
	RoBERTa	0.3439	0.2715	0.2804	(2e-5, 8, 15)
	DistilBERT	0.2856	0.2532	0.2484	(2e-5, 16, 10)
5	BERT	0.3726	0.2910	0.2974	(2e-5, 8, 15)
	RoBERTa	0.3183	0.2625	0.2816	(3e-5, 16, 15)
	DistilBERT	0.2761	0.2454	0.2500	(3e-5, 16, 10)

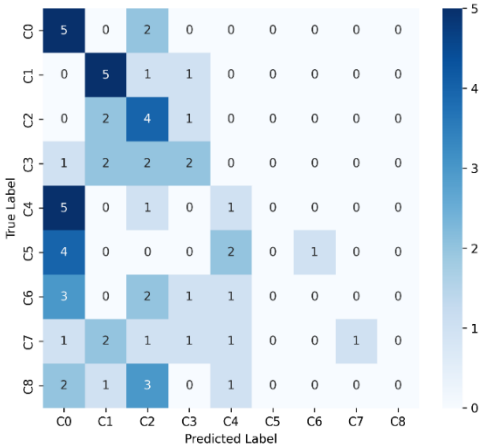

Figure 3. Confusion matrix for best test result using data 1 (DistilBERT)

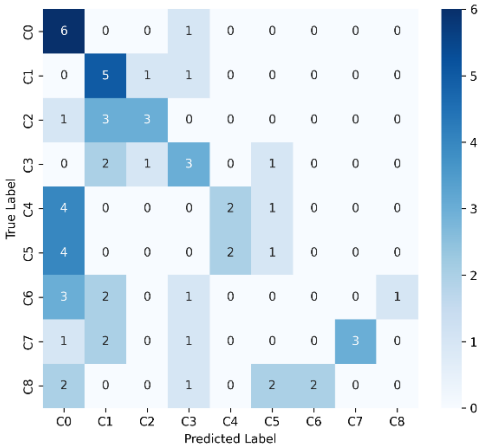

Figure 4. Confusion matrix for best test result using data2 (RoBERTa)

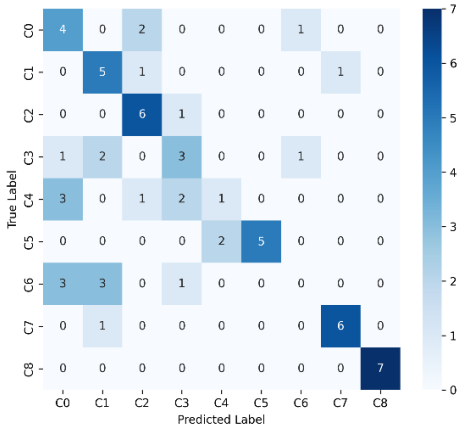

Figure 5. Confusion matrix for best test result using data 3 (DistilBERT)

5 Limitations and Future Work.

Different sentence-level augmentation techniques can be further investigated. Using large language models like Open AI API [23] for generating similar text could also be a potential research methodology. This study focused solely on identifying responsible personnel, but its scope can be expanded to explore other aspects of the crane accident dataset, such as predicting accident-related damages in terms of property and equipment and analyzing root causes of the accidents.

6 Conclusion

In this study, we addressed the challenge of class imbalance in a 9-class classification task for identifying responsible personnel in crane accidents by leveraging paraphrasing techniques for data augmentation. The results across Data 1, 2, and 3 demonstrate that increasing the dataset size through augmentation significantly improved performance. The F1 score of the best-performing model increased from 0.29 (Data 1) to 0.3261 (Data 2) and further to 0.5583 (Data 3).

Through our experiments, we evaluated the effectiveness of paraphrased data in improving the performance of transformer-based models, showcasing the potential utility of widely used paraphrasing tools.

However, this paraphrase-based data augmentation approach has certain limitations. The results may vary depending on the specific settings and methods used to generate the augmented dataset. For future studies, alternative methodologies, such as utilizing large language models for data augmentation, could be explored. Furthermore, experimenting with different hyperparameter settings may yield additional performance improvements.

References

[1] J. D. Wiethorn. An analytical study of critical factors of lift planning to improve crane safety based on forensic case studies of crane accidents. 2018. On-line: http://hdl.handle.net/2152/67528, Accessed: Nov. 13, 2024.

[2] Y. Ji and F. Leite. Automated tower crane planning: leveraging 4-dimensional BIM and rule-based checking. *Automation in Construction*, vol. 93, pp. 78–90, 2018.

[3] Y. Fang, Y. K. Cho, and J. Chen. A framework for real-time pro-active safety assistance for mobile crane lifting operations. *Automation in Construction*, vol. 72, pp. 367–379, Dec. 2016.

[4] V. Herrera-Pérez, F. Salguero-Caparrós, M. del C. Pardo-Ferreira, and J. C. Rubio-Romero. Key Factors in Crane-Related Occupational Accidents in the Spanish Construction Industry (2012 –2021). On-line: https://www.mdpi.com/1660-4601/20/22/7080, Accessed: Mar. 13, 2025.

[5] Y. Suh. Sectoral patterns of accident process for occupational safety using narrative texts of OSHA database. *Safety Science*, vol. 142, p. 105363, 2021.

[6] J. A. Taylor, A. V. Lacovara, G. S. Smith, R. Pandian, and M. Lehto. Near-miss narratives from the fire service: A Bayesian analysis. *Accident Analysis & Prevention*, vol. 62, pp. 119–129, 2014.

[7] J. M. Graves, J. M. Whitehill, B. E. Hagel, and F. P. Rivara. Making the most of injury surveillance data: Using narrative text to identify exposure information in case-control studies. *Injury*, vol. 46, no. 5, pp. 891–897, 2015.

[8] Z. Ma and Z.-S. Chen. Mining construction accident reports via unsupervised NLP and Accimap for systemic risk analysis. *Automation in Construction*, vol. 161, p. 105343, 2024.

[9] L. Kumi, J. Jeong, and J. Jeong. Data-driven automatic classification model for construction accident cases using natural language processing with hyperparameter tuning. *Automation in Construction*, vol. 164, p. 105458, 2024.

[10] J. Chen, D. Tam, C. Raffel, M. Bansal, and D. Yang. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 191–211, 2023.

[11] J. He, J. Gu, J. Shen, and M. Ranzato. Revisiting Self-Training for Neural Sequence Generation. *arXiv*, 2020. arXiv:1909.13788.

[12] J. Chen, Z. Yang, and D. Yang. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, 2020.

[13] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv*, 2020. arXiv:1912.08777.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 2019. arXiv:1810.04805.

[15] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The Curious Case of Neural Text Degeneration. 2020, *arXiv*: arXiv:1904.09751.

[16] M. Peeperkorn, T. Kouwenhoven, D. Brown, and A. Jordanous. Is Temperature the Creativity Parameter of Large Language Models? *arXiv*, 2024. arXiv:2405.00492.

[17] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311-318, Philadelphia, Pennsylvania, 2001.

[18] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating Text Generation with BERT. *arXiv*, 2020. arXiv:1904.09675.

[19] R. Bawden, B. Zhang, L. Yankovskaya, A. Tättar, and M. Post. A Study in Improving BLEU Reference Coverage with Diverse Automatic Paraphrasing. *arXiv*, 2020. arXiv:2004.14989.

[20] Y. Liu *et al.* RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv*, 2019. arXiv:1907.11692.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv*, 2020. arXiv:1910.01108.

[22] Hugging face. Auto Classes. On-line: https://huggingface.co/docs/transformers/en/model_doc/auto, Accessed: Jan. 15, 2025.

[23] OpenAI. OpenAI API. On-line: https://openai.com/index/openai-api/, Accessed: 03/13/2025.