Quality Assessment of Open-Source Language Model Generated Sentence Pairs in Rotorcraft Aviation Domain*

Emma L. McDaniel
Computer Science Department
Georgia State University
Atlanta, GA, USA
emcdaniel10@gsu.edu

Alicia I. Ruvinsky
US Army Engineer Research
and Development Center,
Vicksburg, MS, USA
alicia.i.ruvinsky@erdc.dren.mil

Abstract

We investigate the quality of generated sentence pairs from an open-source language model to be utilized downstream to fine-tune a sentence transformers within the rotorcraft aviation domain. The generated sentences are evaluated using various measurements for prompt adherence, word usage, syntactic similarity, semantic similarity, and style.

1 Introduction

Unstructured data–such as narratives or natural language reports–can be mined and repurposed for a broader range of contexts (Boulton and Hammersley, 2006). In the domain of mining aviation accident and incident reports, many employ topic mining to identify key themes and patterns (Luo and Shi, 2019; Rose et al., 2020) and classification to categorize records based on their content (Zhang et al., 2021; Madeira et al., 2021; Miyamoto et al., 2022). Recent research highlights the effectiveness of language models fine-tuned on domain-specific data (Kierszbaum et al., 2022; Chandra et al., 2023; Jonk et al., 2023).

Mining information deficiencies from aviation accident and incident reports holds potential for creating an information deficiency landscape. We leverage a dataset comprising 8,500+ rotorcraft specific records from the National Transportation Safety Board (NTSB) and National Aeronautics and Space Administration's Aviation Safety Reporting System (NASA ASRS). By linking deficiencies with other flight attributes (e.g., phase of flight, weather), we aim to build an information deficiency landscape to be incorporated in a decision support system for rotorcraft pilots.

This extended abstract explores the initial steps of generating and validating custom sentence pairs tailored for fine-tuning a sentence transformer (Reimers and Gurevych, 2019). Given the potential for "hallucinations" in generative models (Maynez et al., 2020; Ji et al., 2023), a robust validation strategy is crucial to ensure their effectiveness of fine-tuning using these generated sentence pairs.

Table 1: Seven prompts utilized to generate sentences

#	Prompt
1	Rewrite this sentence:
2	Rewrite this sentence to be more informal:
3	Rewrite this sentence to be more formal:
4	Reorganize this sentence:
5	Rewrite this sentence to be more concise:
6	Rewrite this in another way:
7	Rewrite this sentence using different vocabulary:

2 Methodology

We evaluated sentence pairs generated by the Instruct variant of Mixtral (8x7B SMoE), an open-source language model (INSTMixtral; Jiang et al., 2024). We tested 7 "rewrite" prompts, see Table 1, on 50 sampled sentences from our dataset.

To evaluate the quality of the 350 generated sentences, we compared them to the original sentences utilizing a variety of metrics. First, to evaluate word usage similarity and syntactic similarity, we utilized the stemmed form of words (NLTK) and part of speech tags (spaCy) in the sentences in two ways: 1) We calculated the Stem Union % and POS Union %, which represents the percentage of stemmed words/part of speech tags shared between sentences divided by the number of stemmed words in the original sentence; 2) We measured the Stem Lev Dist % and POS Lev Dist % using a normalized Levenshtein distances (LevLibrary) accounting for sentence length, to quantify the similarity

* This research was supported in part by an appointment to the Department of Defense (DOD) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the DOD. ORISE is managed by ORAU under DOE contract number DE-SC0014664. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of DOD, DOE, or ORAU/ORISE.

Table 2: Median scores calculated across prompts (excluding Flesch-Kincaid, Flesh, and Extra %). "Average" is the mean of median scores per prompt. "OS" stands for original sentence score in readability metrics.

Metrics	Average	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7
Stem Union %	63.05%	71.42%	48.53%	64.17%	91.67%	71.43%	60.77%	33.33%
Stem Lev Dist %	80.67%	78.47%	81.82%	82.09%	73.03%	63.64%	89.17%	96.43%
POS Union %	99.2%	100%	100%	100%	100%	94.38%	100%	100%
POS Lev Dist %	57.02%	72.25%	50%	50%	64.06%	50%	62.83%	50%
Flesch-Kincaid	*OS:12.14*	13.61	7.93	16.3	13.01	10.57	13.61	14.65
Flesch	*OS:45.2*	38.08	71.4	25.34	40.87	49.75	39.37	29.14
Cosine Sim	0.836	0.869	0.782	0.823	0.907	0.833	0.859	0.777
Extra %	12%	8%	28%	18%	8%	14%	2%	6%

in stemmed word and tag order. Then, we assessed semantic similarity by embedding the sentences using the All-MPNet-base-v2 (MPNet) and retrieving its cosine similarity score (SK) between its original sentence and the generated sentence. To analyze stylistic differences between the generated and original sentences, we calculated Flesch Reading Ease and Flesch-Kincaid Grade Level (RD) scores for each group of sentences (originals, prompts, etc.). In order to do this, we aggregated each set into a single text to better represent overall stylistic trends within each group. Finally, we assessed prompt adherence, counting the occurrences of extraneous information introduced by the generated versions. The code and dataset utilized in this work are available on our open science framework repository: `https://osf.io/9rtfy/`.

3 Results and Discussion

To assess the quality of the generated sentences, we calculated metrics as they relate to lexical, semantic, syntactic, and stylistic similarity as well as prompt adherence. The results of these metrics are shown in Table 2. For each prompt, we calculated the median value (instead of average) because we found that there were outliers.

Utilizing the union and Levenshtein distance for stemmed words and part of speech tags allowed for assessing similarity in relation to lexical and syntactic usage between the original and generated sentences. The stemmed terms metrics indicates that Prompt 4 has the highest level (91.67%) of similarity, while Prompt 7 has the least similarity (33.33%). Interestingly, the average difference in stemmed word order across prompts was high (80.67%), indicating frequent word rearrangements even with similar vocabulary. Notably, Prompt 5 required the least editing in word order (63.64%), while Prompt 7's sentences would require the most restructuring (96.43%). Part of speech tags, due to the limited set of tags, exhibited high similarity between original and generated sentences. However, the ordering differed with an average of 57%; this suggests that despite similar tags, the structures varied between the original and the generated sentences.

To analyze stylistic differences, we calculated Flesch-Kincaid Grade Level and Flesch Reading Ease scores. These readability scores revealed significant variations across generated prompts. Prompt 2 had the lowest grade level score, but the highest readability ease. This aligns with the prompt's objective of generating informal sentences.

Assessing semantic similarity proved challenging with the current approach. Cosine similarity on generated All-MPNet-base-v2 (MPNet) embeddings to original sentences produced low values despite perceived semantic closeness observed in manual evaluation. Therefore, it is necessary to explore alternative methods to evaluate semantic similarity in the future.

Across all the generated sentences, 12% had extraneous results in the prompt, with Prompt 2 having the highest at 28%. These extraneous information included conversations and extra sentences.

4 Conclusion

Our initial analysis reveals distinct differences in vocabulary, syntax, and style between generated and original sentences. However, assessing semantic similarity necessitates further investigation. This research lays the groundwork for validating generated sentence similarity/dissimilarity for use in fine-tuning a sentence transformer using previously non-paired sentences.

Limitations

This is a preliminary study with a small sample size of 50 records. This raises concerns about the gen-

eralizability of findings to the entire dataset. Additionally, the model's capability to generate novel information not explicitly present in the original sentence presents a potential challenge. Without a more robust method to check for semantic similarity, this could be a limitation of this methodology.

Ethics Statement

Language Models have been shown to reflect and amplify societal biases present in training data. This is an ongoing area of research, with efforts focused on mitigating bias in generated content. While we acknowledge the potential for bias in our generated sentence pairs, the domain of our dataset makes the occurrence of harmful bias less likely.

References

David Boulton and Martyn Hammersley. 2006. Analysis of unstructured data. *Data collection and analysis*, 2:243–259.

Chetan Chandra, Xiao Jing, Mayank V Bendarkar, Kshitij Sawant, Lidya Elias, Michelle Kirby, and Dimitri N Mavris. 2023. Aviation-BERT: A Preliminary Aviation-Specific Natural Language Model. In *AIAA AVIATION 2023 Forum*, page 3436.

INSTMixtral. Mixtral-8x7B-Instruct-v0.1. Version: 4.36.2, Docs: https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Patrick Jonk, Vincent de Vries, Rombout Wever, Georgios Sidiropoulos, and Evangelos Kanoulas. 2023. Natural language processing of aviation occurrence reports for safety management. In *Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)*, pages 2015–2023, Dublin, Ireland.

Samuel Kierszbaum, Thierry Klein, and Laurent Lapasset. 2022. ASRS-CMFS vs. RoBERTa: Comparing Two Pre-Trained Language Models to Predict Anomalies in Aviation Occurrence Reports with a Low Volume of In-Domain Data Available. *Aerospace*, 9(10):591.

LevLibrary. Levenshtein Python C Extension Module. Version: 0.24.0, Docs: https://github.com/rapidfuzz/Levenshtein.

Yonghui Luo and Hongwei Shi. 2019. Using lda2vec topic modeling to identify latent topics in aviation safety reports. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, pages 518–523. IEEE.

Tomás Madeira, Rui Melício, Duarte Valério, and Luis Santos. 2021. Machine learning and natural language processing for prediction of human factors in aviation incident reports. *Aerospace*, 8(2):47.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Ayaka Miyamoto, Mayank V Bendarkar, and Dimitri N Mavris. 2022. Natural language processing of aviation safety reports to identify inefficient operational patterns. *Aerospace*, 9(8):450.

MPNet. sentence-transformers/all-mpnet-base-v2. Version: 4.36.2, Docs: https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

NASA ASRS. NASA Aviation Safety Reporting System (ASRS). Date Downloaded: 26 June 2023, Online: https://akama.arc.nasa.gov/ASRSDBOnline/QueryWizard_Filter.aspx.

NLTK. Natural Language Toolkit. Version: 3.7.2, Docs: https://www.nltk.org/api/nltk.stem.html.

NTSB. National Transportation Safety Board (NTSB). Date Downloaded: 20 June 2023, Online: https://data.ntsb.gov/avdata.

RD. Readability metrics. Version: 1.4.5, Docs: https://pypi.org/project/py-readability-metrics/.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rodrigo L Rose, Tejas G Puranik, and Dimitri N Mavris. 2020. Natural language processing based method for clustering and analysis of aviation safety narratives. *Aerospace*, 7(10):143.

SK. Scikit-Learn. Version: 1.4.0, Docs: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html.

spaCy. Version: 3.7.2, Docs: https://spacy.io/usage/linguistic-features.

Xiaoge Zhang, Prabhakar Srinivasan, and Sankaran Mahadevan. 2021. Sequential deep learning from ntsb reports for aviation safety prognosis. *Safety science*, 142:105390.