Univ. of Hildesheim at AuTexTification 2023: Detection of Automatically Generated Texts

Tatjana Scheibe[1], Thomas Mandl[1,*]

[1]*Information Science, University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany*

Abstract

Text generation models can pose a challenge for the legitimacy and authenticity of texts. Large pre-trained models have reached a high level of quality already. This paper presents experiments on classifying whether a text was written by a human or generated by a language model. The paper describes experiments within shared task AuTexTification: Automated Text Identification 2023. The approach is based on a pre-trained model. We selected the DeBERTaV2 model. Our run reached an Macro-F1 score of 67.2 and was ranked on position 15 out of 76 submissions for subtask 1. The paper also presents an analysis of both text classes based on text metrics. The observation of various readability metrics shows that the generated texts tend to show less diversity than human texts.

Keywords

Text classification, Transformer, Chat-GPT, Readability, Evaluation

1. Introduction

Text generation tools have become extremely powerful and there is a great need for the identification of generated text. Therefore, classifiers for making the distinction between text that was written by humans and text which was generated by machines need to be developed and evaluated. The shared task AuTexTification provides a testbed for such research [1, 2]. In an experiment with this task, we developed a classifier based on a large pre-trained model in order to analyze the capabilities of current large and geenrative language models. This paper also intends to analyze the training dataset by quantifying the quality of the text based on readability metrics [3] as well as other lexical metrics. A transparent analysis could be useful for the explainability of text classifiers and for supporting the task of detecting unethical use of language generation.

2. State of the Art

Large language models in NLP like BERT [4] and GPT-3 [5] have reached an elevated level of quality in text analysis and text generation [6]. Algorithms succeed in writing not only

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

scheibe@uni-hildesheim.de (T. Scheibe); mandl@uni-hildesheim.de (T. Mandl)

https://www.uni-hildesheim.de/fb3/institute/iwist/mitglieder/mandl/ (T. Mandl)

0000-0002-8398-9699 (T. Mandl)

sentences, but whole articles, including modifying the writing style. The best-performing systems are currently based on transformers which process a sentence as a sequence of words which can consider context between all words simultaneously [7]. Systems like BERT and GPT-3 complement this basic idea by more complex techniques. BERT is trained to reconstruct masked tokens within a sentence [4]. It can be applied to generate a sentence embedding which can be used for next sentence prediction.

These powerful tools will have consequences for several domains including literature [8], scientific writing [9] and many other professional activities [10]. Online tools e.g., https://quillbot.com/, https://transformer.huggingface.co, https://philosopherai.com enable citizens to work with and experiment AI technology, but also illustrate the limits of even the most up-to-date systems. Thus, these systems succeed in producing grammatically well-formed texts, but display weaknesses when it comes to coherence [11].

Writing is a fundamental component of academic success in educational contexts. Writing is central to our social identities and we are often evaluated by our control of it. The release of ChatGPT (https://openai.com/blog/chatgpt) marks a turn in writing processes in its diverse forms, as AI now influences writing to a considerably higher extent than previous technologies. This will dramatically change our cultural practice(s) of writing [12]. Consequently, the understanding of what it means to write, revise and post-edit is challenged. The use of AI writing to augment human writing skills will have procedural, ethical and pedagogical ramifications that are currently being debated in the media and in various contexts [13]. Particularly within higher education, concerns have been raised about the potential impact of AI-based writing on academic integrity, authorship recognition and critical source analysis. There is also raising concern that AI writing tools could cause societal issues [14] e.g. due to the spread of misinformation [15].

Although there are great opportunities for a proper didactic use of language models [16, 17] there are worries about inappropriate use in academia [18].

The identification of authorship became an important topic. Can text classification technology or can humans reliably detect machine-generated content? Since tools have become very powerful and widely used, the identification of computer-generated text is more and more relevant. Researchers have observed an increase in automatically generated content even in scientific venues [19].

Text classification experiments for distinguishing between human an machine generated content have been promoted for several years [20]. E.g. within the Bot Profiling task in 2019 a F1 score of 0.96 was obtained [21]. Similar values above 0.9 were obtained for several architecture on another collection, however, the authors admit that the systems are vulnerable for adversarial attacks [22]. It has also been pointed out that the level of performance drops below 0.9 when the texts are shorter than 64 characters [23]. Some first collections exist. Some are specific for domains like misinformation [22] and scientific publications [24].

Also OpenAI itself published a classifier which should identify generated text, however, the company admits that it does not work perfectly well [25]. In an experiment with humans, automatically generated reviews were perceived to be as fluent as human-written ones [26]. The cues which humans and computers might be quite different [27]. There are suggestions on how to pursue a test for humans because the methodological setup can influence the outcome [28]. For example, in one study, humans were asked to detect the boundary between human

Table 1
Frequency of the classes in the training set

Class	Percentage	Number of texts
Human	50.36%	17046
Generated	49.64%	16799

and generated text within a document and performed badly [29].

Despite the research available, more studies are necessary to analyse the differences between human-written and machine-generated text if there are any. For example, it is claimed that automatically generated text uses common phrases more often [30]. The shared task AuTexTification [1] contributes to finding the best technologies which can classify successfully. Furthermore, for assessing the features of texts and the quality of language generation, there is a need for further metrics [31].

In our study, we dedicate some effort to obtaining text metrics for both classes in the given dataset. Such an analysis could reveal differences between the two text classes.

3. System Overview and Experiment

Within the shared task AuTexTification, we submitted one run (team Stiftungsuni_Hildesheim) for subtask 1 in English. We intended to solve the task with a pre-trained model in order to judge its quality for such a classification task. For our experiment, we applied the DeBERTaV2 model. For fine-tuning purposes with the traning set, we used the AUTOTRAIN service by Hugging Face to load the models and run training and evaluation sets. The best performance was achieved by the DeBERTaV2 model on a text classification task.

The system was fine-tuned with 3000 randomly chosen texts from the AuTEXTification train dataset. We applied a reduced set in order to keep the computational load low.

On the training set, we obtained the following performance values:

- Accuracy: 0.936
- Precision: 0.922
- Recall: 0.952
- AUC: 0.982
- F1: 0.937

In the result ranking on the test data, the approach reached a Macro-F1 score of 67.2. This drop suggests that the training adopted the system too strongly to the training set features. It is likely, that our model did not perform well for the cross-domain generalization which was the objective of the task [1].

4. Analysis and Discussion

This section reports on our analysis of the text features of the training set. We included several text metrics like readability metrics [32].

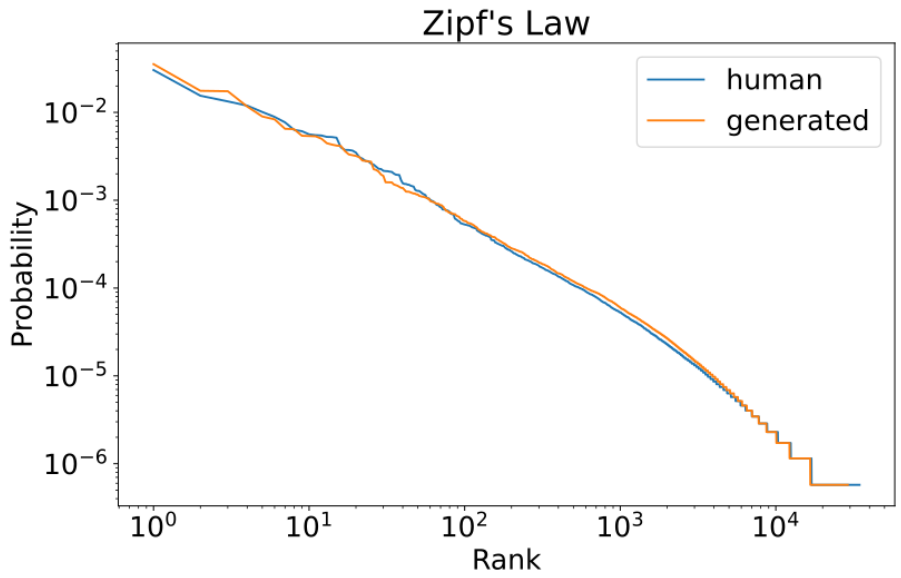

Figure 1: Comparison of the probability distributions of words for the two classes in the training set

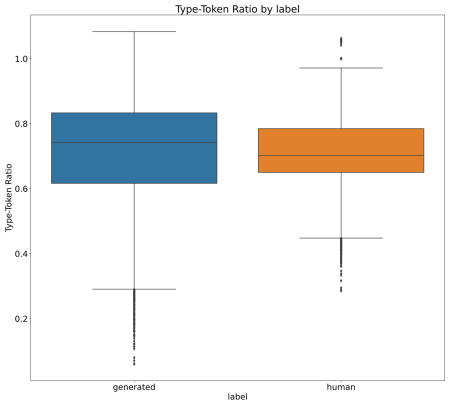

Figure 2: Boxplot for TTR

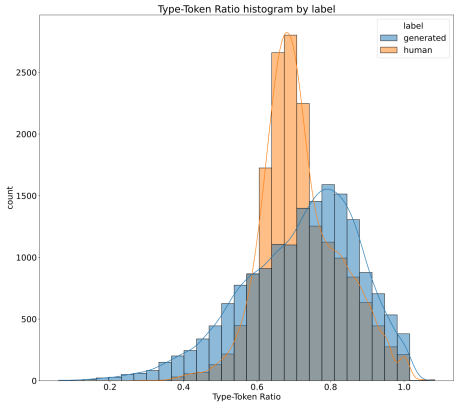

Figure 3: Histogram of TTR

The length of the texts in the training dataset ranges from 1 to 115 words. The size in texts and the distribution over the classes is shown in table 1.

The probability distribution of words follow Zipf's Law in large corpora. There is no deviation in the collection of generated texts which could indicate that the generative model does not create language like humans. Figure 1 shows that the generated texts represent a perfect Zipf distribution. Previous work confirmed this finding [33].

Lexical diversity has also been considered as a key indicator of text quality [34]. It is often

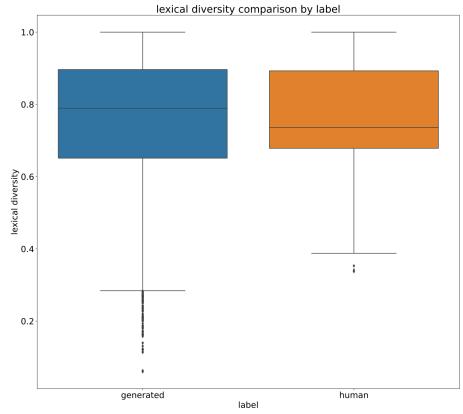

Figure 4: Lexical diversity

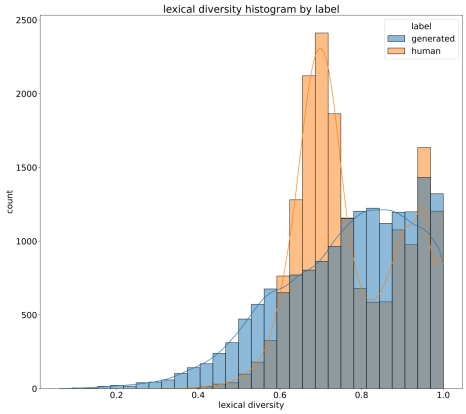

Figure 5: Histogram of Lexical diversity

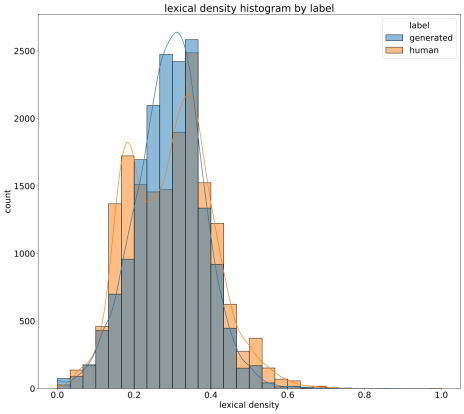

Figure 6: Lexical Density

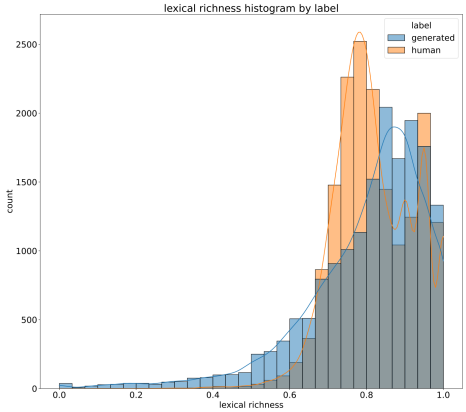

Figure 7: Lexical Richness

used as a synonym to lexical richness or diversity. It is also used to assess human writing [35]. For evaluating the text complexity, a measurement often used is lexical density, which is measured by measures such as Type-Token Ratio (TTR) [36].

We measured the ratio between types and tokens. It can be observed that the generated texts cover a wider range whereas the human texts exhibit much higher values in the distribution around the median value. This is illustrated by the boxplot for the distributions in Figure 2 as well as the histogram in Figure 3.

The same is the case for further metrics. We show the lexical diversity in Figure 4 and in Figure 5. Figure 7 shows the lexical richness in both classes and Figure 6 shows the lexical density. A boxplot of these metrics is given in Figure 8.

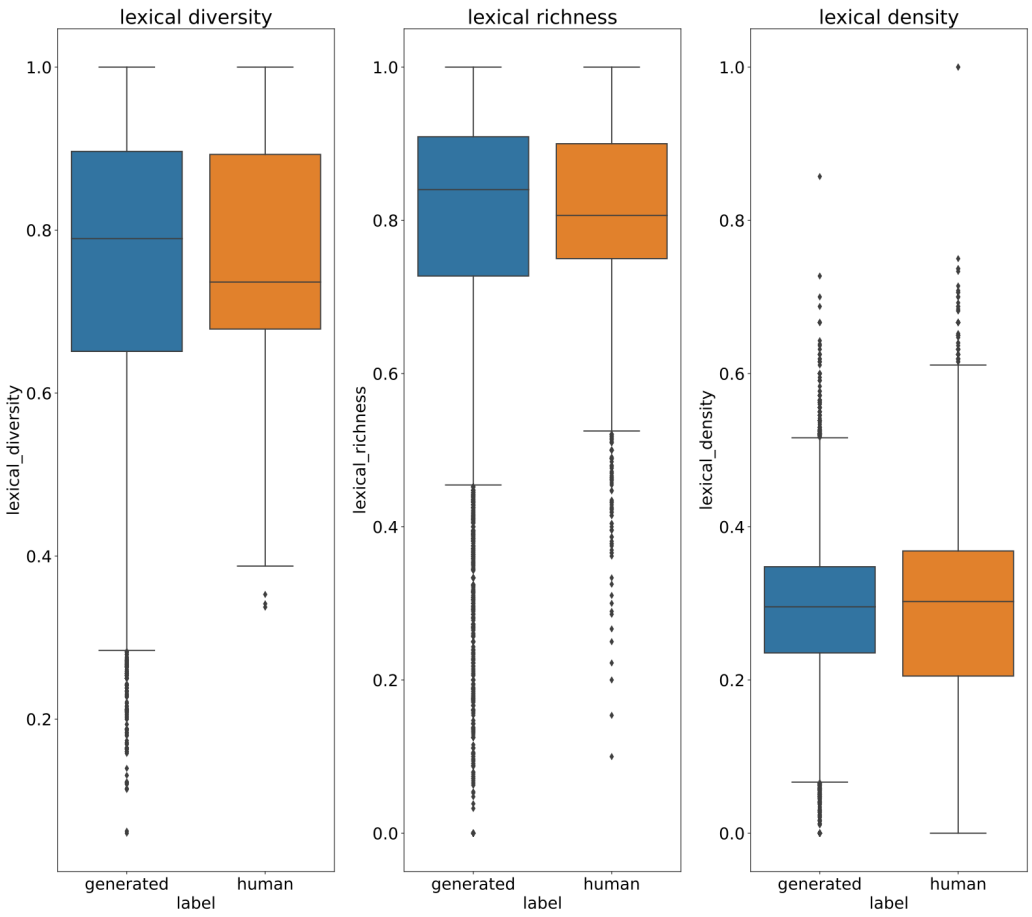

Figure 8: Boxplot of lexical metrics

5. Readability Metrics

Lexical metrics do not provide a full complexity analysis of the text. Readability metrics, like Flesch Reading Ease [37] or Gunning Fox Index [38], are well-known in the US system and some have been used for nearly a century to assess the difficulty of texts in schools. For example, the Flesch Reading Ease measures the complexity of a text and returns values between between 1 and 100. 100 is considered very easy, while 1 is considered very difficult. It was developed by Rudolf Flesch in the 1940s. These metrics focus on the length of texts and the length of words. Recommendations for achieving good scores can be found online (e.g. at http://readable.com).

Figure 9 and Figure 10 show that the generated text in the AuTexTification task does not exhibit different values for the readability metrics. The models generate a range of sentences with varying complexity. However, for some metrics, there seems to be a higher number of texts with a score close to the average. These distributions show a higher peak for values close to the medium when compared to the human generated texts.

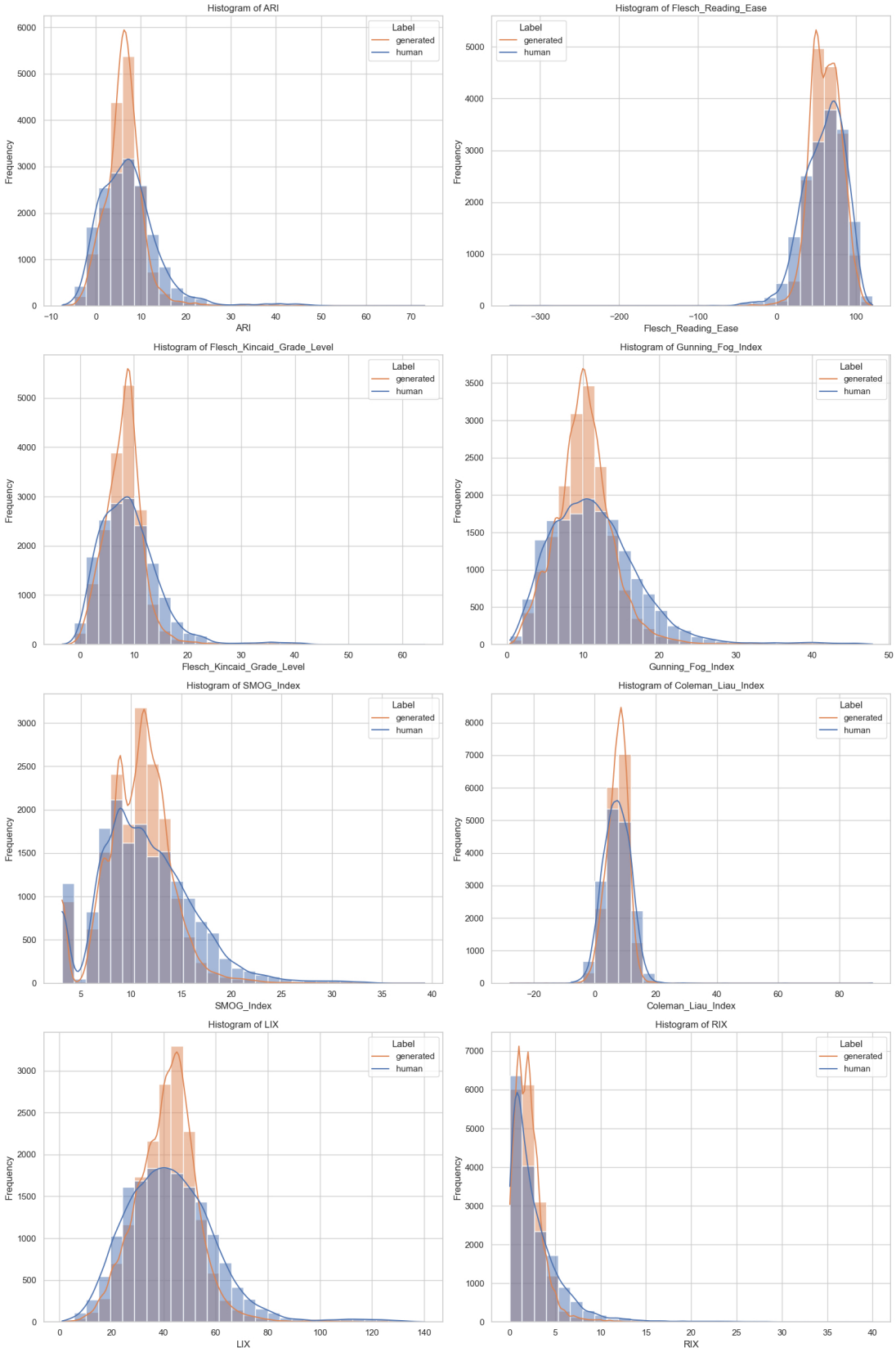

Figure 9: Comparison of several readability metrics for the two classes in the training set

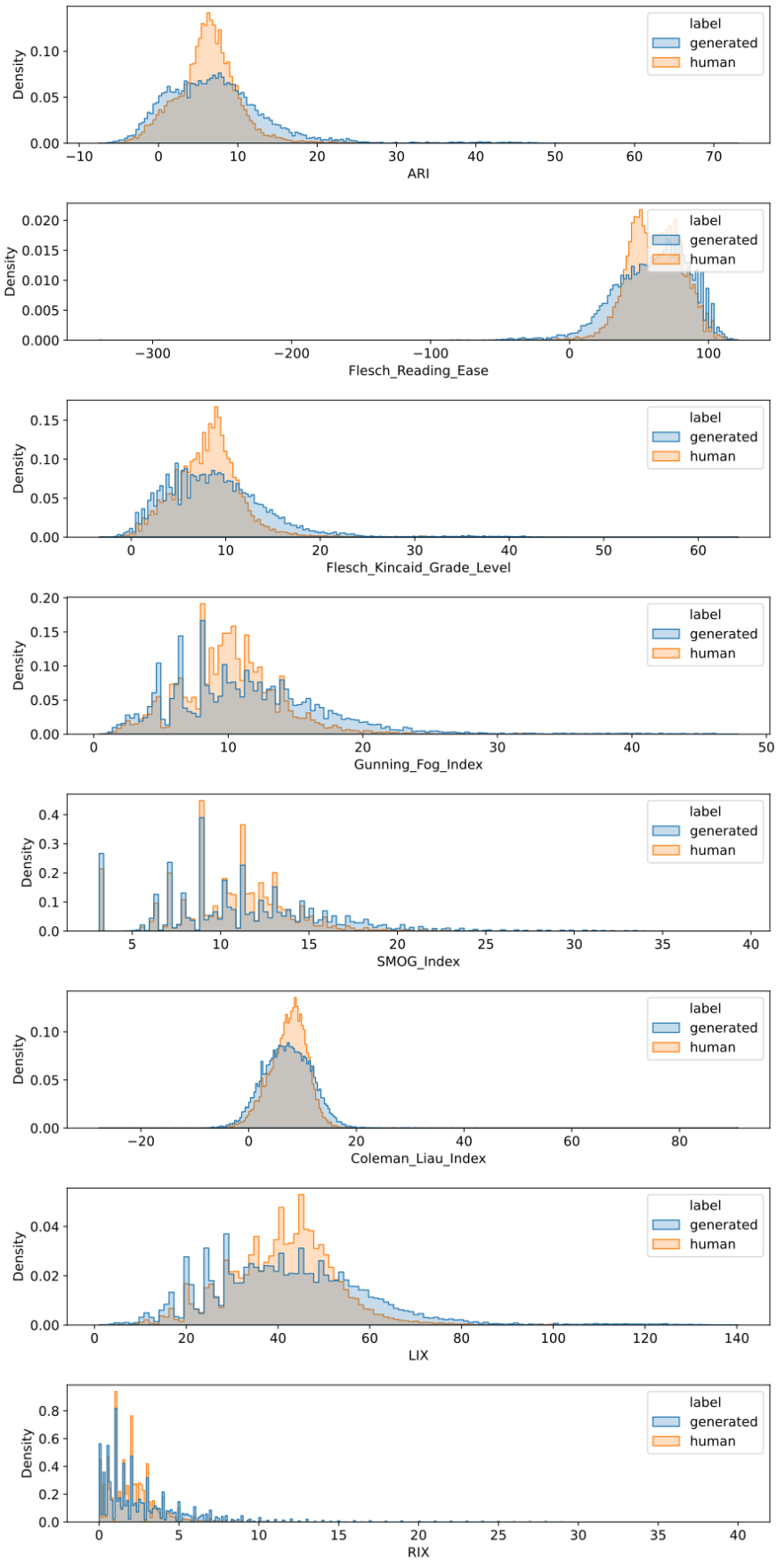

Figure 10: Histogram for several readability metrics for the two classes in the training set

6. Future Work

As future work, we envision to improve our classification system. In addition, further text metrics should be explored. Furthermore, we intend to conduct experiments with humans [28] in order to find out how well humans perform for the domain and the texts selected for AuTexTification.

References

[1] A. M. Sarvazyan, J. Á. González, M. Franco Salvador, F. Rangel, B. Chulvi, P. Rosso, Overview of AuTexTification at IberLEF 2023: Detection and Attribution of Machine-Generated Text in Multiple Domains, in: Procesamiento del Lenguaje Natural, Jaén, Spain, 2023.

[2] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages, Procesamiento del Lenguaje Natural 71 (2023).

[3] E. Pitler, A. Nenkova, Revisiting readability: A unified framework for predicting text quality, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2008, pp. 186–195. URL: https://aclanthology.org/D08-1020.pdf.

[4] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. Conference of the North American Chapter of the ACL: Human Language Technologies, NAACL-HLT, Minneapolis, MN, USA, June 2-7, ACL, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[5] R. Dale, GPT-3: What's it good for?, Natural Language Engineering 27 (2021) 113–118. doi:10.1017/S1351324920000601.

[6] A. Chan, GPT-3 and InstructGPT: technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry, AI Ethics 3 (2023) 53–64. doi:10.1007/s43681-022-00148-6.

[7] S. Modha, P. Majumder, T. Mandl, An empirical evaluation of text representation schemes to filter the social media stream, J. Exp. Theor. Artif. Intell. 34 (2022) 499–525. URL: https://doi.org/10.1080/0952813x.2021.1907792.

[8] A. Elstermann, Computer-generated text as a Posthuman mode of literature production, Open Library of Humanities 6 (2020). doi:10.16995/olh.627.

[9] M. Salvagno, F. Taccone, A. Gerli, Can Artificial Intelligence help for Scientific Writing?, Crit Care 27, 75 (2023). doi:10.1186/s13054-023-04380-2.

[10] E. Felten, M. Raj, R. Seamans, How will language modelers like chatgpt affect occupations and industries?, arXiv preprint arXiv:2303.01157 (2023).

[11] O. Marchenko, O. Radyvonenko, T. Ignatova, P. Titarchuk, D. Zhelezniakov, Improving text generation through introducing coherence metrics, Cybernetics and Systems Analysis 56 (2020) 13–21. doi:10.1007/s10559-020-00216-x.

[12] P. P. Ray, ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems (2023). doi:10.1016/j.iotcps.2023.04.003.

[13] L. Li, Z. Ma, L. Fan, S. Lee, H. Yu, L. Hemphill, ChatGPT in education: A discourse analysis of worries and concerns on social media, arXiv preprint arXiv:2305.02201 (2023).
[14] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, C. Rizzo, ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health, Frontiers in Public Health 11 (2023) 1567. doi:10.3389/fpubh.2023.1166120.
[15] T. Hsu, S. A. Thompson, Disinformation Researchers Raise Alarms About A.I. Chatbots, 2023. URL: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html.
[16] U. Bohle-Jurok, J. Baumgart, T. Mandl, Ki-basiertes Textfeedback in englischsprachigen Lehrveranstaltungen (KI-TextengL), in: TextFeedBack in Praxis und Forschung: 3. gemeinsame Tagung der gefsus, der GeWissS und des Forum Schreiben. 7. - 9. Sept., online., 2023.
[17] J. M. Gayed, M. K. J. Carlon, A. M. Oriola, J. S. Cross, Exploring an AI-based writing Assistant's impact on English language learners, Computers and Education: Artificial Intelligence 3 (2022) 100055. doi:10.1016/j.caeai.2022.100055.
[18] M. Liebrenz, R. Schleifer, A. Buadze, D. Bhugra, A. Smith, Generating scholarly content with ChatGPT: ethical challenges for medical publishing, The Lancet Digital Health 5 (2023) e105–e106. doi:10.1016/S2589-7500(23)00019-5.
[19] B. A. Sabel, E. Knaack, G. Gigerenzer, M. Bilc, Fake publications in biomedical science: Red-flagging method indicates mass production, medRxiv (2023). doi:10.1101/2023.05.06.23289563.
[20] M. S. Aljabri, R. Zagrouba, A. Shaahid, F. Alnasser, A. Saleh, D. M. Alomari, Machine learning-based social media bot detection: a comprehensive literature review, Soc. Netw. Anal. Min. 13 (2023) 20. URL: https://doi.org/10.1007/s13278-022-01020-5. doi:10.1007/s13278-022-01020-5.
[21] F. M. R. Pardo, P. Rosso, Overview of the 7th Author Profiling Task at PAN 2019: Bots and gender profiling in twitter, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, Sept. 9-12, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: https://ceur-ws.org/Vol-2380/paper_263.pdf.
[22] H. Stiff, F. Johansson, Detecting computer-generated disinformation, International Journal of Data Science and Analytics 13 (2022) 363–383. doi:10.1007/s41060-021-00299-5.
[23] A. Pagnoni, M. Graciarena, Y. Tsvetkov, Threat scenarios and best practices to detect neural fake news, in: Proc. 29th Intl. Conference on Computational Linguistics, 2022, pp. 1233–1249. URL: https://aclanthology.org/2022.coling-1.106/.
[24] V. Liyanage, D. Buscaldi, A. Nazarenko, A benchmark corpus for the detection of automatically generated text in academic publications, arXiv preprint arXiv:2202.02013 (2022).
[25] J. H. Kirchner, L. Ahmad, S. Aaronson, J. Leike, New AI classifier for indicating AI-written text, 2023. URL: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.
[26] D. I. Adelani, H. Mai, F. Fang, H. H. Nguyen, J. Yamagishi, I. Echizen, Generating sentiment-preserving fake online reviews using neural language models and their human- and machine-based detection, in: Proc. 34th Intl. Conference on Advanced Information Networking and Applications, AINA, Caserta, Italy, 15-17 April, volume

1151 of *Advances in Intelligent Systems and Computing*, Springer, 2020, pp. 1341–1354. doi:10.1007/978-3-030-44041-1_114.

[27] D. Ippolito, D. Duckworth, C. Callison-Burch, D. Eck, Automatic detection of generated text is easiest when humans are fooled, in: Proc. 58th Annual Meeting of the Association for Computational Linguistics, ACL, Online, July 5-10, 2020, pp. 1808–1822. doi:10.18653/v1/2020.acl-main.164.

[28] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text, in: Proc. 12th Intl. Conference on Natural Language Generation, Association for Computational Linguistics, Tokyo, Japan, 2019, pp. 355–368. doi:10.18653/v1/W19-8643.

[29] L. Dugan, D. Ippolito, A. Kirubarajan, C. Callison-Burch, RoFT: A tool for evaluating human detection of machine-generated text, in: Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations, ACL, Online, 2020, pp. 189–196. doi:10.18653/v1/2020.emnlp-demos.25.

[30] S. Gehrmann, H. Strobelt, A. Rush, GLTR: Statistical detection and visualization of generated text, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL, Florence, Italy, 2019, pp. 111–116. doi:10.18653/v1/P19-3019.

[31] J. Novikova, O. Dušek, A. Cercas Curry, V. Rieser, Why we need new evaluation metrics for NLG, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 2241–2252. doi:10.18653/v1/D17-1238.

[32] M. Martinc, S. Pollak, M. Robnik-Šikonja, Supervised and unsupervised neural approaches to text readability, Computational Linguistics 47 (2021) 141–179. doi:10.1162/coli_a_00398.

[33] S. Chugh, R. Rohilla, Empirical laws of natural language processing for neural language generated text, in: Information, Communication and Computing Technology: 6th International Conference, ICICCT, New Delhi, India, May 8, Revised Selected Papers 6, Springer, 2021, pp. 184–197. doi:10.1007/978-3-030-88378-2_15.

[34] Y. Wang, J. Deng, A. Sun, X. Meng, Perplexity from PLM Is Unreliable for Evaluating Text Quality, arXiv preprint arXiv:2210.05892 (2022).

[35] J. Read, Assessing Vocabulary, Cambridge University Press, 2000.

[36] N. Kapusta, M. Müller, M. Schauf, I. Siem, S. Dipper, Assessing the Linguistic Complexity of German Abitur Texts from 1963–2013, in: Proceedings of the 18th Conference on Natural Language Processing (KONVENS), 2022, pp. 48–62. URL: https://aclanthology.org/2022.konvens-1.7.pdf.

[37] R. Flesch, How to write plain English, 1979. URL: https://web.archive.org/web/20160712094308/http://www.mang.canterbury.ac.nz/writing_guide/writing/flesch.shtml.

[38] O. S. Goh, C. C. Fung, A. Depickere, K. W. Wong, Using Gunnnig-Fog index to assess instant messages readability from ECAs, in: Third International Conference on Natural Computation (ICNC), volume 5, IEEE, 2007, pp. 480–486. doi:10.1109/ICNC.2007.800.