IDENTIFYING TEXT CLASSIFICATION FAILURES IN MULTILINGUAL AI-GENERATED CONTENT

Raghav Subramaniam

Independent Researcher

ABSTRACT

With the rising popularity of generative AI tools, the nature of apparent classification failures by AI content detection softwares, especially between different languages, must be further observed. This paper aims to do this through testing OpenAI's "AI Text Classifier" on a set of human and AI-generated texts inEnglish, German, Arabic, Hindi, Chinese, and Swahili. Given the unreliability of existing tools for detection of AI-generated text, it is notable that specific types of classification failures often persist in slightly different ways when various languages are observed: misclassification of human-written content as "AI-generated" and vice versa may occur more frequently in specific language content than others. Our findings indicate that false negative labelings are more likely to occur in English, whereas false positives are more likely to occur in Hindi and Arabic. There was an observed tendency for other languages to not be confidently labeled at all.

KEYWORDS

Artificial Intelligence, AI Detection, Generative AI, GPT, Natural Language Processing

1. INTRODUCTION

Artificial Intelligence (AI) has become notable for its applications in a wide range of demonstrated capabilities, a popular one being AI-generated text. With the advent of widely available and easily-accessible language models, such as GPT-3.5, many members of the public are able to use AI models (with relative ease) capable of producing highly sophisticated multilingual text [1]. While these capabilities bring certain benefits, it also gives area for concern regarding the potential misuse of text generated by AI for unethical purposes (a concern that has been widely expressed in the past) [2]. The problem this paper aims to center around is the risk of the blanket word "cheating" which can be done through usage of AI-generated content (if falsely claimed to be original work).

The problem of cheating brings the need for tools that are able to detect AI-generated text. Such tools have been available, however, their accuracy on multiple languages (especially non-English languages) may need further examination. Considering the development of language models like GPT-3.5 (*ChatGPT* is built on this) which many detection tools are trained on, English has primarily been the language in which training data is selected. When AI detection tools are trained to detect text generated by these models, this English-centric input could result in a potentially significant bias and lack of necessary empirical training when applied to non-English languages [3]. This bias presents a challenge when it comes to accurately detecting AI-generated text in non-English input. This study aims to examine the extent of the gap (if there exists a gap) in the performance and effectiveness of AI text detection systems when identifying non-English AI-generated content when compared to English AI-generated content. Due to the lack of publicly available information on these detection tools, it is difficult to fully understand their

biases without conducting very large numbers of test cases, however by conducting a comprehensive evaluation of the various mentioned styles and languages of input,we seek to better understand what AI detection vulnerabilities arise when observing certain languages. This can hopefully aid in addressing possible solutions as well.

Existing work on this topic will first be addressed in Section 2. In Section 3, our methods utilizing a collected corpus of both AI-generated and human-written content (consisting of 96 texts across genres and languages) will be discussed along with analysis of results.

2. RELATED WORK

It has been established that it is difficult for humans to identify AI-generated text in English alone, as one case example shows how in an incentivized version of the Turing Test, participants repeatedly failed to differentiate between AI-generated poetry, and that of real poets, meaning we cannot rely on humans to make this differentiation on their own [4]. While this is a significant milestone in AI development, the societal implications in multiple fields are still important to consider, especially as governments and communities across the world consider the ethical and moral implications of AI, one of the most commonly considered possibilities being in schooling: cheating on assignments/assessments. As a result of easily-accessible generative AI models, it becomes possible for students to now use prompts to generate school assignments. This can become a problem when students claim such AI-generated pieces of text to be their own. While it was already a problem with humans not being able to differentiate between AI-generated content and human-created content, the same problem appears to persist even when turning to AI tools to make this differentiation. In one example in the United Kingdom, when prompting AI to complete an essay writing assessment from an accredited university's physics module, responses were estimated to earn "First Class", the highest classification available at UK universities. When examining this with AI-generated text classifiers, however, responses scored only 7% AI-generated on TurnItIn's detection service and only 2% AI-generated on Grammarly's detection service [5]. False predictions like this could result in many university students graduating through possible gaps in knowledge, leading to more problematic implications: such negative reverberations of cheating in schooling systems have always been present, however, they can potentially be exacerbated to a very large extent by this rising risk of undetected AI-generated text.

This observation in English-language environments (UK universities), also persists when foreign-language environments are observed, which actually contain differently-natured errors in attempts to identify AI-generated texts [6]. In Wee & Reimer's research, when human-writtentext in Malay, Mandarin, and Japanese were given to 3 AI detectors, they were all repeatedly incorrectly flagged as AI-generated when they were in fact written by humans. This provides another risk, as honest students in university or schooling environments may incorrectly be assumed to be cheating based on the false predictions of such detectors. In academia this problem may become alarming as well, as researchers may have their work incorrectly accused of being AI-generated. Rather than following a format similar to these studies which identify this general problem, our research seeks to specifically address the nature of such accuracy differences between various languages and how certain biases may become apparent (all done in the context of Open AI's AI Text Classifier).

OpenAI's text classifier is not the only option however, as previous research on improving AI tasks in non-English languages has been completed. These tools however are not as widely implemented for this specific use case of AI-generated text detection. One example is XLM-RoBERTa, a model pre-trained on 2.5 TB of filtered Common Crawl data containing 100 languages that combines two popular approaches in natural language processing: XLM (Cross-

lingual Language Model) and RoBERTa, which can be used for a variety of multilingual classification tasks, including identification of AI-generated text [7]. This can potentially be leveraged for its breadth of multilingual capabilities to create a fine-tuned model for multilingual AI content detection, but is beyond the scope of this paper

3. PERFORMANCE ANALYSIS

In order to achieve this, we will utilize "ChatGPT", a tool trained on the GPT-3.5 model, to generate text in languages with a sizable population of native speakers that are majority non-European and from various parts of the world.

The aforementioned model will be used to generate text in 6 languages: Swahili, German, Arabic, Chinese (simplified), Hindi, and English. To ensure that a single type of generated text is not over-represented, we will use a dataset of AI-generated pieces of text with a range of genres including stories, poems, informational articles, and argumentative pieces. 8 pieces of text will be generated for each language as follows: two stories, two poems, two informational articles, and two argumentative pieces. Topics and positions will be randomly picked. They willbe generated with the following prompt "Write a (story / article / poem / argumentative piece) about (topic and position if applicable) in 1200 characters". Each of these 48 generated pieces of content will be accompanied with a same-language same-topic public domain digital writing piece for reference, which will have similar length. These human-written pieces will be pulled from Project Gutenberg and the Internet Archive. This provides a total of 48 AI-generated pieces, and 48 human-generated pieces, all 96 of which will then be inputted into Open AI's *AI Text Classifier*. Once these results are labeled on a scale of "very unlikely AI-generated" to "likely AI-generated" (labeling practices for the purposes of this paper include abbreviations discussed in Section 3.1) we analyze the results in Section 3.2.

3.1. Labeling Results

It is important to note that Open AI's publicly available AI Text Classifier assigns texts to one of 5 possible labels to complete a grammatically correct variation of the sentence "The classifier considers the text to be [label] AI-generated". In this context, we have assigned each label to 5 different abbreviations: "UC" stands for "unclear", "P" stands for "possibly" (possibly AI-generated), "L" stands for "likely" (likely AI-generated), "UL" stands for "unlikely" (unlikely to be AI-generated), and "VU" stands for "very unlikely" (very unlikely to be AI-generated).

As the experimental methods were carried out, certain patterns in distribution of the aforementioned labels were evident. Table 1 is an example of these initially observed distribution of labelings for each language (Swahili, German, Arabic, Simplified Chinese, Hindi, and English), expressed as a proportion of the amalgamated corpus of text generated by both artificial intelligence and human agents

Table 1. Labelings as a percentage (decimal representation) of the combined AI/Humangenerated text corpus

Language	Percentage of total "P" labelings	Percentage of total "UL" labelings	Percentageof total "L" labelings	Percentageof total "UC" labelings	Percentageof total "VU" labelings
English	0.03	0.41	0.00	0.04	0.50
Hindi	0.40	0.06	0.00	0.11	0.00
Arabic	0.30	0.00	0.88	0.00	0.00
Swahili	0.00	0.18	0.00	0.37	0.21
German	0.13	0.18	0.13	0.26	0.07
Chinese (Simplified)	0.13	0.18	0.00	0.22	0.21

Despite small variances it can be seen at large that there is a trend of certain languages being affected with more uncertainty in classification than others, specifically when considering the "UC" label. Languages like English and Arabic tended to have the lowest share of "UC" labels (unclear whether the text was AI-generated), however languages like Swahili, German, and Chinese had fairly significant shares. This leads to a reasonable assumption that more definite predictions with more confidence would be attributed to these same languages: English and Arabic. When observing the more definite predictions of "VU", very unlikely, and "L", likely, English and Arabic not only have the largest share of these labels, but they also have almost a majority of these labels as a whole. English has around 50% of the "VU" labels while Arabic accounts for 88% of the "L" labels.

An interesting point to consider in these two languages is that while English had a large percentage of "VU" labels, Arabic had no share of these labels: the same pattern exists vice versa for the "L" label. While Arabic had a significant share of "L" labels, English had no share. This roughly correlates with the previously mentioned study conducted in UK Universities, in which a very significant majority of high-scoring AI-generated physics essays were not detectedby the AI-content classifier [5]. This could possibly point to bias in OpenAI's classification tool that prefers for English text to be classified as unlikely and very unlikely the majority of the time (a point to consider especially as English also has the highest percentage of "UL" labelings). There may be specific engineering goals that lead to this outcome, for example, on OpenAI's landing page for AI Text Classifier it is stated that "we adjust the confidence threshold to keep the false positive rate low; in other words, we only mark text as likely AI-written if the classifier is very confident" [8]. Higher exposure to English texts during training may make the classifier more likely to achieve this goal in this language, leaving other languages like Arabic, on the other end of the spectrum. This also means that AI-generated text used for cheating could more easily go undetected if it were in English (as opposed to Arabic). As the vast majority of "L" labelings were given to Arabic input, it's quite possible that Arabic could have more false positives for AI-generated text detection. If true, this would roughly mirror results from the study conducted by Wee and Raimer, in which inputs in 3 non-English (East Asian and Southeast Asian) languages were mislabeled as AI-generated by many classifier tools [6]. Although their study incorporated more AI classifier tools than Open AI's tool, a similar phenomenon in training data may have persisted for the other classifiers as well.

3.2. Distribution Of False Labelings

In Table 2, percentages of total false positives and false negatives for each language are provided False positives refer to human-written text misclassified as "P" or "L", while false negatives refer to AI-generated text misclassified as "UL" or "VU".

Table 2. False labelings per language as a decimal percentage of each category (Generated byHuman/AI)

Language	**Percentage of Total False Positives**	**Percentage of Total False Negatives**
English	0.06	0.88
Hindi	0.31	0.00
Arabic	0.50	0.00
Swahili	0.00	0.00
German	0.13	0.00
Chinese (Simplified)	0.00	0.11

The previous connections made to both the UK university study and the East/Southeast Asian language study therefore hold relatively true as we observe that English accounted for a significantly lower percentage of false positives then languages like Hindi and Arabic (accounting for 31% and 50% respectively), which together accounted for 81% of false positive labelings. This is in stark contrast to the results from English, which turn out to follow Open AI's original disclaimer that their AI classifier tool only marks text as AI-generated if there is high confidence: English appears to account for almost none of the false positives (6%) while at the same time accounting for 88% of total false negatives. As a result of this tendency to mark English text as not AI-generated, real cheating in schooling (in English) may easily go undetected, whereas the case with non-English languages may not be the same. This correlates with the previously presented example of AI-generated English text going undetected [5] while some non-English human-written texts (in this case Hindi and Arabic) are more frequently mislabeled as AI-generated when using the same tool. This is a pattern that was observed in other studies with non-English languages as well [6], and could possibly indicate that the AI text detection tool discussed could lead to incorrect assumptions of cheating when examining text in these languages: a problem opposite to the one created in English text.

This leaves the question of why Swahili, German, and Chinese also had noticeably low rates of false positives. This could possibly be due using the results from Table 1, which show that these 3 languages accounted for a large proportion of "UC" labelings: combined, 84% of these labelings originated from this group of languages alone, which is presumably why the proportion of false positives and false negatives are also low for this group. The apparent inability of AI Text Classifier to make confident predictions in this group is interesting as some non-English languages like Arabic, which accounted for 88% of total "L" labelings, were able to (although erroneously) be confidently labeled more frequently than other non-English languages. Especially as it has been mentioned in previous literature that creation of Arabic generative AI presents unique challenges [9], it is notable how instead of classifying Arabic texts as ambiguously as Swahili, German, and Chinese, the detection tool has more labeling confidence. When considering this, a possible vulnerability of this study is how all texts were restricted to a length of 1200 characters. Certain languages like Arabic have an average word length of 5 characters, while languages like Chinese have an average word length of 2 characters. This could lead to a larger number of words in a 1200-character Chinese text than in a 1200-character Arabic text. This may have possibly triggered a bias in the AI text detector to have more confidence with more/fewer words, which could have possibly been fixed (if it were to be a significant problem) by using a word target instead of a character target. Regardless, these results still point to areas in which unique biases are apparent and could be helpful in answering the question of how these AI text detection tools can be improved to help more accurately identify AI-enabled cheating in not just English, but multiple global languages as well.

3.3. Distribution of Positives and Negatives by Genre

Texts were also split by genre with two types of creative writing (stories and poems) and two types of formal writing (informational articles and argumentative pieces). Distributions among the 4 genres are examined below.

Table 3. Majority Positive Languages by Genre (Generated by Human/AI)

Genre	Languages with majority positivesin AI-generated text	Languages with majority positives inhuman-written text
Stories	HIN, ARB	HIN, ARB
Poems	HIN, ARB, GER, CHN	ARB
InformationalArticles	HIN, ARB	ARB
ArgumentativePieces	ARB	HIN, ARB, GER

In Table 3, languages for which the majority of ratings were positive are distributed by genre and source. In the context of this paper, "positive" refers to ratings of "L" or "P", both referring to a relatively higher level of certainty (than other ratings) that the text is AI-generated. The table also employs abbreviations for each language. "HIN" refers to Hindi, "ARB" refers to Arabic, "GER" refers to German, "CHN" refers to Chinese, and "SWA" refers to Swahili. As can be seen in the table, the majority of these majority positive labelings come from the languages identified to have a large proportion of total "L" and "P" labelings in Table 1. The continuation regardless of genre is notable as it is shown that the phenomenon of the two languages (Hindi and Arabic) being frequently marked as AI-generated expands over the lines of creative/formal writing. Although one may assume intuitively that a single type of writing may be more frequent in training data (such as formal over creative) there seems to be no difference. Another possibility to consider could be that the AI detection service itself is not capable of semantic "understanding" of the text that is being presented to it in the same way thatit may be able to with other languages such as English. This could explain why content in certain languages is consistently marked as AI-generated or unclear. Variations in this phenomenon such as the absence of Hindi in some of the boxes above may be attributed to othernon-semantic reasons such as text length, sentence size, etc.

Table 4. Majority Negative Languages by Genre (Generated by Human/AI)

Genre	Languages with majority negatives in AI-generated text	Languages with majority negatives inhuman-written text
Stories	ENG	ENG
Poems	ENG	SWA, CHN
InformationalArticles	ENG	ENG, SWA, GER
ArgumentativePieces	ENG	ENG

In Table 4, languages for which the majority of ratings were negative are distributed by genre and source. In the context of this paper, "negative" refers to ratings as "VU" or "UL", both referring to a low level of likelihood that the text is AI-generated (likely that it is written by a human). The same abbreviations for language from Table 3 are also used in Table 4. As can be seen in the table, the majority of these majority negative labelings come exclusively from the English language. As observed in Table 3, the same lack of variance between genres of text persists as creative writing often has the same "majority negative" language as formal writing. A reason for this lack of variance could be attributed to the same disclaimer discussed in Section 3.2: the lack of "positive" confidence across genres may be a reason for these repetitive assumptions of human-written status among English-language texts.

4. CONCLUSION

Throughout our work, we examine certain variations in classification accuracy arising from application of OpenAI's *AI Text Classifier* to multilingual texts across various genres of creative and formal writing. Even with the decision to use a character limit instead of a word limit and our usage of a relatively compact dataset used (instead of a large scale set of multilingual text of various writing genres), the findings can provide some clarity regarding the nature of the various classification failures that arise with both false negative labelings (as demonstrated in English) and false positive labelings (as demonstrated in Hindi and Arabic). As described in past work on the subject [4, 5], the implications are large in a variety of fields, especially in schooling systems in which classification failures can severely impact understanding of "fair" work in certain languages. Accordingly, attention must be brought to the current state of AI-generated text classification failures present in English and non-English languages alike. This is imperative for working towards enhancing AI-text classification tool performance across all languages and genres. Although exploring the linguistic features that lead to failures in each language is beyond the current scope of this paper, this could be further examined in future investigations. Future studies could also use a word-based limit as opposed to a character-based one, as well as use a wider range of writing styles and languages in order to reveal reasons behind persistent errors within particular language families or writing styles: these are factors that could assist in creation of higher accuracy AI classification tools. Until this happens, it becomes imperative for communities to remain cognizant about the risks posed by misclassified text. Raising awareness about these risks is crucial in avoiding incorrect allegations of AI-text usage and equally importantly in avoiding unnoticed use of AI-generated text.

REFERENCES

[1] A. B. Mbakwe, I. Lourentzou, L. A. Celi, O. J. Mechanic, and A. Dagan, "ChatGPT passing USMLE shines a spotlight on the flaws of medical education," PLOS Digital Health, vol. 2, no. 2, p. e0000205, Feb. 2023, doi: 10.1371/journal.pdig.0000205.

[2] Y. K. Dwivedi et al., "'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," International Journal of Information Management, vol. 71, p. 102642, Aug. 2023, doi: 10.1016/j.ijinfomgt.2023.102642.

[3] A. Conneau, "Cross-lingual Language Model Pretraining," 2019. https://proceedings.neurips.cc/paper_files/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html

[4] N. C. Köbis and L. Mossink, "Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry," Computers in Human Behavior, vol. 114, p. 106553, Jan. 2021, doi: 10.1016/j.chb.2020.106553.

[5] W. Yeadon, O.-O. Inyang, A. Mizouri, A. Peach, and C. P. Testrow, "The death of the short-form physics essay in the coming AI revolution," Physics Education, vol. 58, no. 3, p. 035027, Apr. 2023, doi: 10.1088/1361-6552/acc5cf.

[6] H. B. Wee and J. D. Reimer, "Non-English academics face inequality via AI-generated essays and countermeasure tools," BioScience, Apr. 2023, doi: 10.1093/biosci/biad034.

[7] A. Conneau, "Unsupervised Cross-lingual Representation Learning at Scale," arXiv.org, Nov. 05, 2019. https://arxiv.org/abs/1911.02116

[8] "New AI classifier for indicating AI-written text." https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text (accessed Jul. 03, 2023).

[9] S. Alhumoud, A. A. Wazrah, and W. Aldamegh, "Arabic Chatbots: A Survey," International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, Jan. 2018, doi: 10.14569/ijacsa.2018.090867.