JOURNAL OF
EMERGING INVESTIGATORS

Article

Part of speech distributions for Grimm versus artificially generated fairy tales

Suvrath Arvind[1], Clayton Greenberg[2]

[1] Prospect High School, Saratoga, California

[2] Saarland University, Graduate School of Computer Science, Saarbrücken, Saarland, Germany

SUMMARY

ChatGPT is a chatbot tool that relies on GPT3 and later OpenAI transformer language models to generate responses to user prompts. In this study, we sought to investigate the statistical differences between naturally generated and artificially generated text due to the dramatic increase in quality of natural language generation from large language models, popularized by ChatGPT. To constrain our problem, we considered fairy tales as these texts have existed for centuries. To explore statistical differences, we focused on the distribution of words according to their parts of speech (POS), elements that characterize words based on their grammatical function. We generated a novel corpus of 101 fairy tales "authored" by ChatGPT. We compared this against 209 fairy tales written by the Grimm Brothers and made available freely online. Our hypothesis was that the distributions of POS for Grimm fairy tales and ChatGPT fairy tales would be different and that the POS distributions will vary among Grimm fairy tales more than among ChatGPT fairy tales. We performed appropriate preprocessing and computed total variation distances for individual fairy tales within and between authorship conditions. We found out that in fact, the distribution of POS in ChatGPT fairy tales is significantly different from the distribution of POS in Grimm fairy tales.

INTRODUCTION

Released in late 2022, ChatGPT quickly became one of the most downloaded apps due to its accessibility and ability to quickly generate natural language responses to various prompts. We wanted to explore if this holds for fairy tales, a foundational part of many human cultures. ChatGPT is quite new, having only been released ten months before this research was conducted, and since it received a lot of attention following its release, we wanted to investigate if ChatGPT responses had some of the same statistical properties as human-written text.

About two centuries ago, the Grimm Brothers compiled arguably one of the most famous sets of fairy tales, with their own distinct style. These fairy tales are read around the world to this day; however, the use of artificial intelligence has made us wonder if it is possible for a machine to create fairy tales similar to those created by the Grimm Brothers. In order to see if that is the case, we can compare the types of words in each set of fairy tales, and see if the distributions are similar or significantly different. Parts of speech (POS) are classifications of words based on the grammatical purpose they serve in a sentence. For instance, the subject of a sentence is always a noun, noun phrase, or pronoun, while the action of the sentence is always a verb or verb phrase. In the simple sentence, "Cats run," the word "cats" would be a noun as it is the subject, while the word "run" would be a verb since it is the action. More complex POS would include adjectives, adverbs, and interjections. For an effective comparison, we used an English translation of the original Grimm fairy tales, written in German.

Every word in the dictionary and every proper noun has a POS associated with it, but unlike the thousands of words that exist in the English language, there are only a handful of POS. Comparing words using POS, rather than the actual words themselves, allows us to make generalizations over smaller amounts of texts and accurately explore more of the differences between fairy tales created by two vastly different authors. In every story, there can be subtle differences, including in character and setting names. To effectively compare the general ideas and the ways words are used in a fairy tale (or any story), POS are used, as these subtle differences do not impact POS distributions. In a way, POS distributions can be more flexible than distributions of the actual words themselves. By using a large number of artificially generated fairy tales (in our case, 101 fairy tales), we can be confident that there would be enough of each POS to get a good estimate of their relative proportions, but we would not be as accurate in estimating the proportion of many specific words in the text. This is because it is not computationally feasible with the current interface for ChatGPT (an interactive user request-system response tool) to collect reliable statistics at the individual word level. ChatGPT, unlike authors, does not use a consistent style when creating fairy tales. Fairy tales are created by humans to discuss cultural aspects of various time periods. However, ChatGPT, due to its algorithm, does not have a specific culture, so the topics it creates fairy tales about may not be as consistent as human authors. In short, humans create fairy tales with a reason; ChatGPT creates fairy tales because a user tells it to do so.

Similar research has been conducted on comparing ChatGPT generated text with that produced by humans, including articles and essays (1, 2). However, there has not been any research done on comparing fairy tales, one of the fundamental building blocks of storytelling, generated by artificial intelligence tools with fairy tales generated by human authors. Storytelling is an essential part of human culture and civilization, and comparing complex pieces of writing, such as fairy tales, would allow us to understand if artificial intelligence can compose stories in ways similar to humans. Also, other research works have not analyzed POS distributions, but rather analyzed how humans perceive artificially generated texts or how accurate grammatical techniques are used in

https://doi.org/10.59720/23-326

artificially generated texts. By conducting this research, we will be able to understand how similar or different ChatGPT is from humans. Finding differences in POS distributions could aid developers of artificial intelligence models in evaluating texts created by those programs and the differences could possibly suggest ways to improve the models. In addition, we wanted to explore the variability in POS distributions between each set of fairy tales (Grimm and ChatGPT) to determine if there is a consistent way of writing in each set of fairy tales.

Since ChatGPT produces responses using an algorithm, our hypothesis was that the POS distribution between Grimm fairy tales and ChatGPT fairy tales would be different. Also, since humans wrote Grimm fairy tales, we hypothesize that the POS distributions will vary among Grimm fairy tales more than among ChatGPT fairy tales. Our research showed that our hypotheses were correct, with the POS distributions for Grimm and ChatGPT fairy tales being different, and with Grimm fairy tales having a POS distribution with greater variation than that of ChatGPT. Since artificially generated texts lack the variation present in human produced texts, artificial intelligence detection softwares can use POS distributions to determine whether writing was artificially produced.

RESULTS

After compiling the set of 101 ChatGPT fairy tales using the prompt "tell me a fairy tale," we observed that when it came to proper nouns, many characters' names (and their roles) were repeated multiple times. For example, the character Lily was present in 43 out of the 101 fairy tales and was often the protagonist, while the character Morgana (often "Queen Morgana") was present in 33 fairy tales and was often the antagonist. Just like character names, place names also shared commonalities, such as names that are similar to the word "Enchanted" (including Enchanted Forest and Enchantia), the setting of 44 fairy tales, and places that had names similar to the word "Everland" (including Everland and Everlandia), the setting of 20 fairy tales. We decided to keep all of the 101 original ChatGPT fairy tales despite the similarities between each fairy tale since this set would provide the most opportunity to capture the diversity of ChatGPT responses.

We next used tokenization to compare the ChatGPT-derived and Grimm fairy tales. Tokens are the computational analog of words. The simplest scheme for tokenization (breaking up a text into tokens) splits up the text using white space, which may produce undesirable tokens at punctuation marks and other special symbols. POS tagging assigns tags per token, so the choice of tokenization scheme is pivotal for a POS distribution analysis. Analyses based on text length and (trivially) token distribution are also dependent on the choice of tokenization scheme. We found that on average, each ChatGPT fairy tale had about 3517 characters (letters, spaces, and symbols), which formed about 589 tokens (tokenizing on whitespace). Each token in the entire set of ChatGPT fairy tales had an average length of 6 characters. When tokenizing using the Natural Language Toolkit (NLTK), each fairy tale, on average, had 679 tokens (3). Most of the ChatGPT fairy tales had between 600 and 700 tokens, which is quite different from human-produced fairy tales, which have greater variation, as different authors use different amounts of text to convey their ideas (**Figure 1A**). This is seen by the observation that Grimm fairy tales had about 1561 tokens, unlike ChatGPT's 679 tokens, and that the number of tokens for all Grimm fairy tales was skewed, meaning that there is greater variation in the number of tokens used in the Grimm fairy tales and that the data points are not evenly distributed on either side of the mean (**Figure 1B**). This variation showed that unlike artificially generated texts, human created texts are not based on an algorithm, but rather come from human creativity and experiences.

We then began our analysis of the POS distributions by comparing ChatGPT fairy tales to Grimm fairy tales. The first observation we made was that in the fairy tales created by ChatGPT, nouns, verbs, and adjectives combined to form a large proportion (47.27%) of words (**Figure 2**). In addition, many POS tags had different proportions in each set of fairy tales (**Figure 3**). For example, ChatGPT had nearly double the proportion of nouns (represented by the tag 'N') in its stories, while Grimm fairy tales had more verbs ('V'), punctuation marks ('.'), predeterminers, and pronouns (the

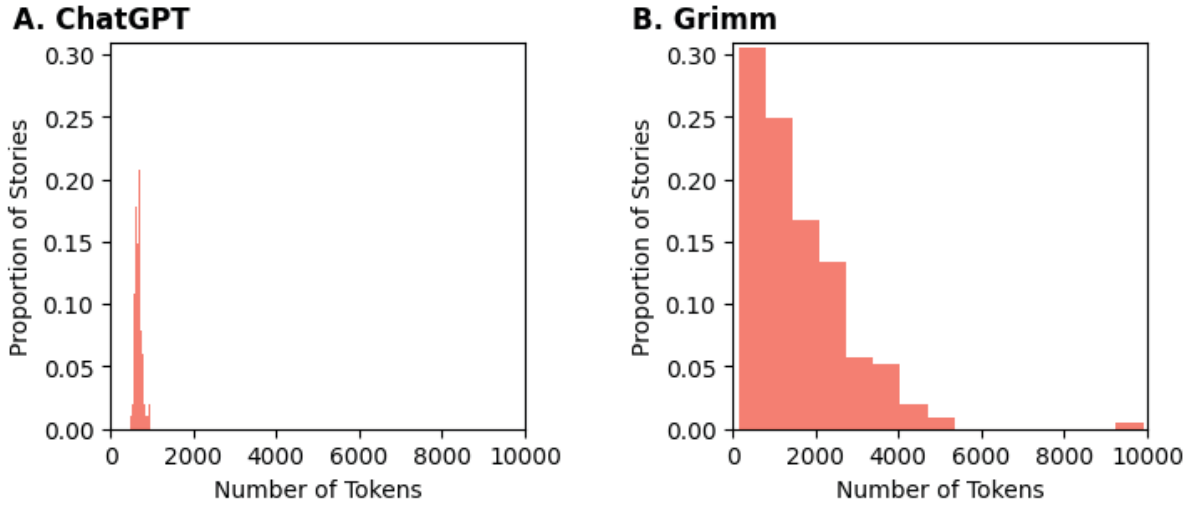

Figure 1: Number of tokens in fairy tales. Each of the ChatGPT fairy tales (A) and Grimm fairy tales (B) were tokenized using the Natural Language Toolkit and the number of tokens for each fairy tale was recorded (n = 101 ChatGPT fairy tales and n = 209 Grimm fairy tales).

https://doi.org/10.59720/23-326

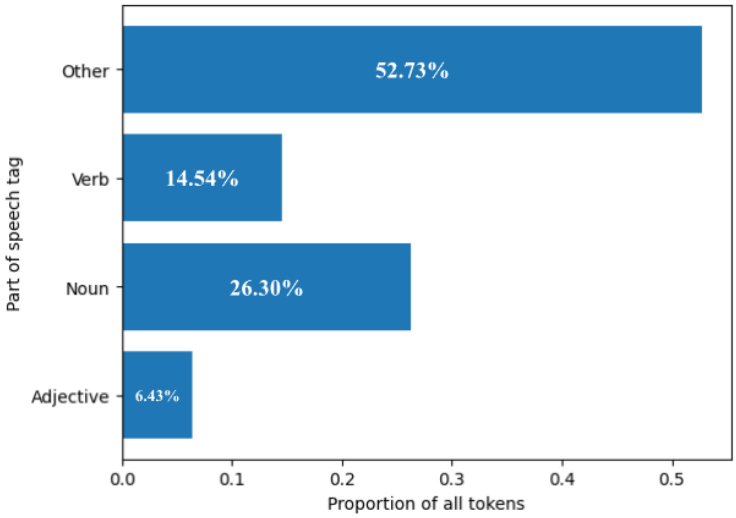

Figure 2: Proportions of certain parts of speech (POS) in ChatGPT fairy tales. Horizontal bar chart showing the proportions of adjectives, nouns, verbs, and other POS tags in ChatGPT fairy tales (n = 101). Each of the ChatGPT fairy tales were tokenized using the Natural Language Toolkit and then all of the tokens were categorized into nouns, verbs, adjectives, and other.

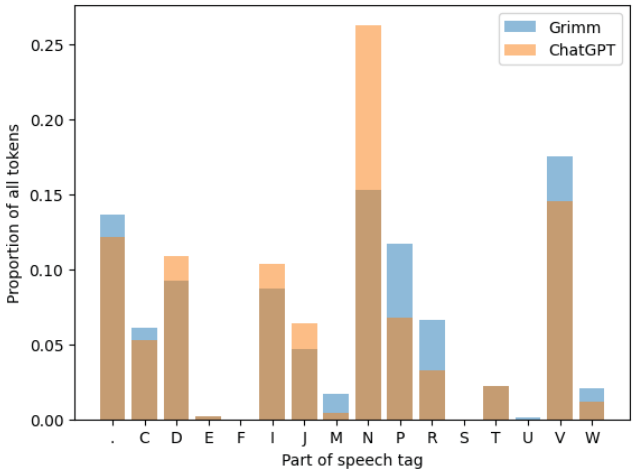

Figure 3: Proportions of simple parts of speech (POS) tags in ChatGPT and Grimm fairy tales. Bar chart showing the proportions of each POS tag in both sets of fairy tales (n = 2). The proportion of each POS tag was compared between ChatGPT and Grimm fairy tales. The definitions of POS tags are as follows: Punctuation Mark (.), Conjunction (C), Determiner (D), Existential (E), Foreign Word (F), Preposition (I), Adjective (J), Modal (M), Noun (N), Predeterminer and Pronoun (P), Adverb (R), Symbol (S), 'To' (T), Interjection (U), Verb (V), Words beginning with 'Wh' (W).

last two both being represented by 'P'). Using this, we found that the calculated Total Variation Distance (TVD), a measure of how far apart (or different) two distributions are from each other, was 0.16. To determine this value's significance (or likeliness of this occurring simply due to chance), we used the proportions of POS tags in the Grimm fairy tales as the probabilities associated with each POS tag in a large number (10,000) randomly generated stories and calculated the TVD between each of the 10,000 stories and the set of Grimm fairy tales to see what would be probable TVDs between a randomly generated story and the Grimm fairy tales. Using that, we were able to determine if our observed TVD was significant (and was unlikely to be captured by the randomly generated TVDs) to see if in fact the POS distributions were significantly different. This approach led us to find that our observed TVD of 0.16 was highly significant (p-value < 0.0001, bootstrapping test, **Figure 4**).

We verified our observed TVD as we also compared each individual ChatGPT fairy tale to the Grimm fairy tales. The distribution of TVDs for each ChatGPT fairy tale against the Grimm fairy tales was approximately normal, with a mean slightly larger than 0.16, a value similar to our observed TVD, further supporting our finding that the two POS distributions are significantly different from each other (**Figure 5**).

We then compared the Grimm fairy tales to themselves, using the leave-one-out cross validation technique (4). We iteratively selected one fairy tale from the set of 209 Grimm fairy tales and then calculated the TVD between the POS distribution of that fairy tale and the POS distribution of the other 208 remaining fairy tales. This allowed us to see how each individual fairy tale compared to the rest of the fairy tales in a given set, which allowed us to see variation within that set. The leave-one-out TVDs for the Grimm fairy tales formed a distribution that was skewed to the right, indicating that there was variation in the writing produced by humans, something that was expected (**Figure 6A**). When we applied the leave-one-out cross validation technique to the ChatGPT fairy tales, we found that the distribution of the leave-one-out TVDs for ChatGPT fairy tales was approximately normal (**Figure 6B**). There is significantly less variation in artificially generated fairy tales compared to human-produced fairy tales (p-value < 0.0001, Bootstrapping Test).

DISCUSSION

From our initial observations, we found that nouns, verbs, and adjectives made up a large proportion of the words in the ChatGPT fairy tales. This was expected since these three POS make up a large proportion of the English language. However, we observed that verbs were the most common POS in Grimm fairy tales while nouns were the most common POS in ChatGPT fairy tales, despite both being the

https://doi.org/10.59720/23-326

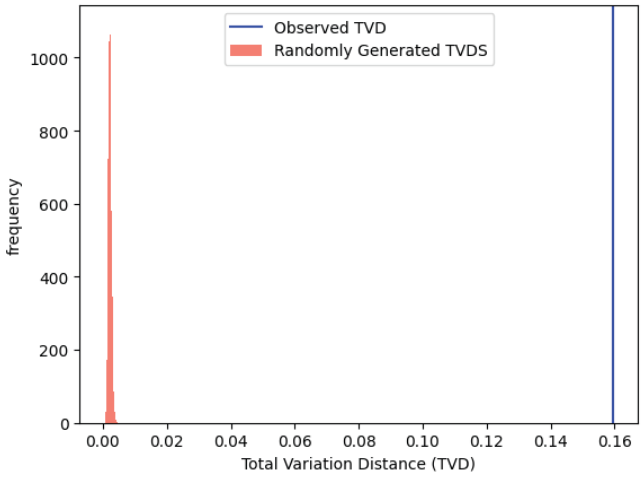

Figure 4: Randomly generated TVDs versus our observed TVD. Histogram showing randomly generated TVDs between ChatGPT and Grimm fairy tales (n = 10000). The TVD (about 0.16) was compared with randomly generated TVDs in order to determine the significance of our result. Bootstrapping Test, *** *p*-value < 0.0001.

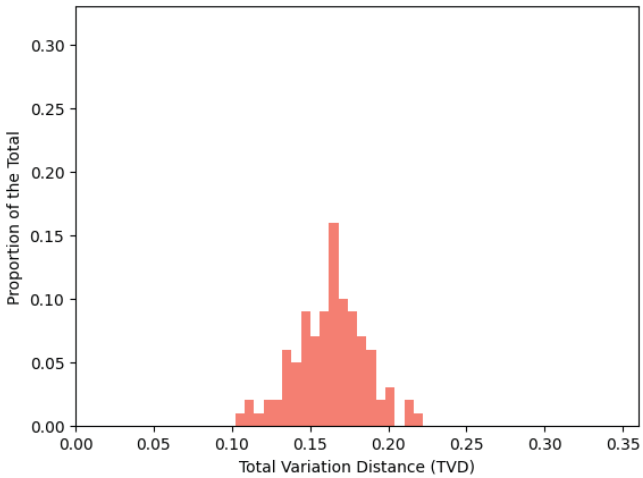

Figure 5: TVDs for ChatGPT fairy tales against Grimm fairy tales. Histogram showing TVDs between each ChatGPT fairy tale and the set of Grimm fairy tales (n = 101). The TVD between each ChatGPT fairy tale and the set of 209 Grimm fairy tales was recorded, and the mean was found to be slightly larger than 0.16.

two most used POS used in each set of fairy tales.

After beginning our analysis, we observed that there was a slight difference between our observed TVD and the mean of the distribution of the TVDs calculated between each ChatGPT fairy tale and the entire set of Grimm fairy tales. This difference can be attributed to the fact that the latter was determined by comparing the POS distribution of each ChatGPT fairy tale individually to the set of Grimm fairy tales instead of finding the total POS distribution of the ChatGPT fairy tales and then comparing that to the POS distribution of the Grimm fairy tales.

In addition, our comparison between ChatGPT and Grimm fairy tales revealed stark differences in their POS distributions due to the significantly large TVD, confirming our hypothesis. This showed that fairy tales generated by artificial authors are not structured in the same way as fairy tales produced by human authors. Our result from the leave-one-out TVDs for the Grimm fairy tales, that there was variation in the POS distributions of each of the Grimm fairy tales, was expected since it is very unlikely for a human to write the same way (using the same number of each POS) every time they create a new piece of writing. In contrast, the POS distribution of each of the ChatGPT fairy tales was approximately normal, which was surprising. We were not sure if ChatGPT would have the same variability as Grimm when producing fairy tales, since we expected ChatGPT to write fairy tales in ways similar to humans. However, this result can be justified since machines, unlike humans, lack the variability in works produced.

In conclusion, we found that despite being used to produce writing similar to that of humans, ChatGPT lacks the variability

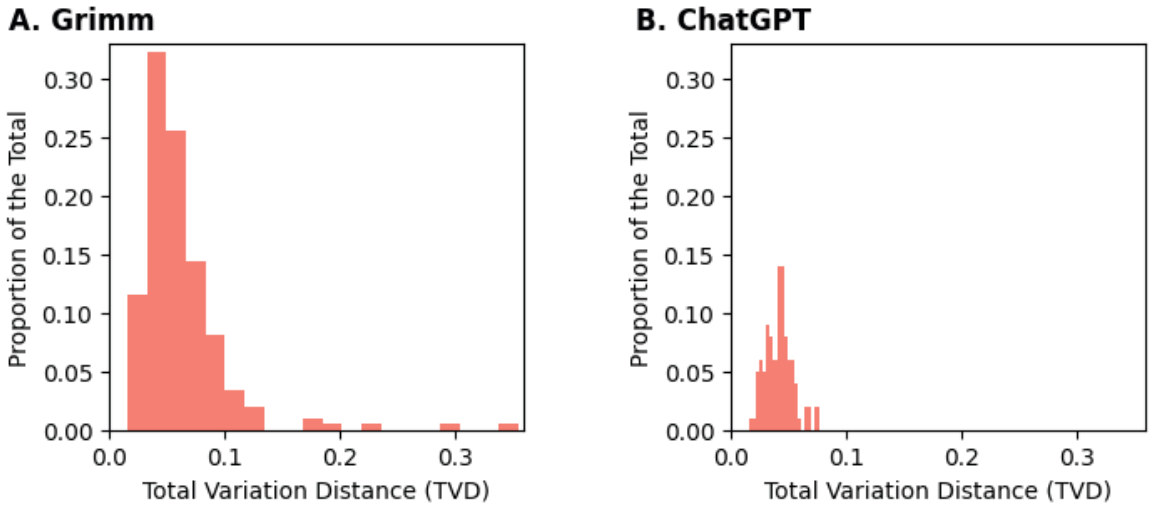

Figure 6: Leave-one-out TVDs for each set of fairy tales. Histogram showing the leave-one-out TVDs calculated for Grimm fairy tales (A) and ChatGPT fairy tales (B) (n = 209 Grimm fairy tales and n = 101 ChatGPT fairy tales). Each Grimm fairy tale was compared to every other Grimm fairy tale systematically, and the TVDs between each fairy tale and the remaining fairy tales were recorded and the same process was repeated for the ChatGPT fairy tales.

https://doi.org/10.59720/23-326

that human writing has, and the text it generates is simply artificial. Machines and tools like ChatGPT were developed to produce perfect work, but what separates a machine or artificial intelligence tool from a human is the perfectness a machine has. In order to be similar to humans, a machine must be trained to be imperfect rather than perfect, and ChatGPT, like many other tools, was simply too perfect. It is important to note that we only compared ChatGPT to Grimm rather than a broader set of authors due to their massive impact on modern fairy tales. Therefore, this result is only applicable to fairy tales created or compiled by the Grimm Brothers and there is a chance that ChatGPT might be similar to other authors. We could expand this research to compare ChatGPT (or other artificial sources) with other fairy tale authors or other types of texts and see if there were differences between other POS distributions.

In addition, finding statistical differences in POS distributions between artificially generated text and human produced text could be used in a classifier to detect the use of artificial intelligence in creating texts. The POS distribution of human written texts that are similar to the text being classified and the POS distribution of the text could be compared to see if there are statistical differences; if there are in fact differences, that text could have been generated artificially. Given the almost too perfect nature of texts generated artificially seen in their POS distributions, human created and artificially created texts can easily be distinguished by comparing their POS distributions.

MATERIALS AND METHODS

In order to create a collection of representative ChatGPT fairy tales, we prompted the tool with the command "tell me a fairy tale." Using this prompt, we created a novel set of 101 fairy tales (a large number to get an accurate sample of possible fairy tales that ChatGPT could produce). We made this set of fairy tales available at our GitHub repository (5). For the Grimm Brothers' fairy tales, we used a compilation of 209-fairy tales from the Carnegie Mellon University Computer Science department (6).

Due to the difference in size between Grimm and ChatGPT fairy tales, we compared differences between the proportions of POS tags between the two sets of fairy tales. The POS tags were generated by the NLTK maximum entropy tagger (3). We were able to tag the 101 ChatGPT fairy tales and 209 Grimm fairy tales using the Penn Treebank tagset (7). However, since the 45 tags that the program generated were likely too much, we decided to simplify it further. Each POS tag was a string of two to three characters, and tags relating to the same general POS started with the same letter. We decided to use only that first letter when tagging all the words, and as a result we had 16 simplified POS tags to work with. We then calculated the proportion of each POS tag in all 101 ChatGPT fairy tales and 209 Grimm fairy tales in order to conduct our analysis.

To compare the two datasets, we used the Total Variation Distance (TVD), a statistic that describes how different two probability distributions are from each other. The lower the TVD, the closer the two distributions are in similarity. To determine whether the observed TVD was significant, we found TVDs between 10,000 randomly generated stories (that had POS distributions where each POS had a probability equal to the respective proportion in the Grimm fairy tales) and the Grimm fairy tales and then comparing that set of TVDs to our observed TVD from the initial comparison of Grimm and ChatGPT fairy tales. Then, using the bootstrapping test, we calculated the probability of our observed TVD happening simply by chance to determine if our TVD was significant.

We also used a leave-one-out analysis technique to compare the distribution of fairy tales in each set to each other. This process takes one randomly selected fairy tale out of its dataset and determines the TVD between that and each of the other fairy tales in that set. After creating the set of all TVDs, we would be able to understand how different each fairy tale is from each other, giving us a better picture of the variation within each set of fairy tales.

ACKNOWLEDGMENTS

We would like to thank Polygence for enabling this project to happen. This platform enabled us to work together to effectively analyze the POS distributions between the two sets of fairy tales.

Received: November 14, 2023
Accepted: April 8, 2024
Published: November 16, 2024

REFERENCES

1. Ariyaratne, Sisith, *et al.* "A Comparison of ChatGPT-Generated Articles with Human-Written Articles." *Skeletal Radiology*, vol. 52, no. 9, 14 Apr. 2023, pp. 1755–1758. https://doi.org/10.1007/s00256-023-04340-5.
2. Fitria, Tira Nur. "Artificial Intelligence (AI) Technology in OpenAI CHATGPT Application: A Review of Chatgpt in Writing English Essay." *ELT Forum: Journal of English Language Teaching*, vol. 12, no. 1, 31 Mar. 2023, pp. 44–58. https://doi.org/10.15294/elt.v12i1.64069.
3. Bird, Steven, *et al.* "Natural Language Processing with Python." www.nltk.org/book. Accessed 30 Sept. 2023.
4. Schneider, Jeff. "Cross Validation." *CMU School of Computer Science*, 7 Feb. 1997. www.cs.cmu.edu/~schneide/tut5/node42.html#. Accessed 18 Feb. 2024.
5. Arvind, Suvrath, and Clayton Greenberg. "Fairy-Tales." *GitHub*, 8 Oct. 2023, github.com/Suvrath-A/fairy-tales. Accessed 8 Oct. 2023.
6. "Grimm's Fairy Tales." *CMU School of Computer Science*. www.cs.cmu.edu/~spok/grimmtmp/. Accessed 30 Sept. 2023.
7. Marcus, Mitchell, *et al.* "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics*, vol. 19, no. 2, 1993, pp. 313–330. aclanthology.org/J93-2004.