Yulia Savine
April 23, 2024

Linguistic Markers Analysis of Tweepfakes: A comparison of human vs. AI-generated Tweets

I. Background

With a rise in AI popularity and accessibility, numerous pushes have been made to distinguish between human vs. AI-generated content, specifically in generated text (Herbold et al, 2023; Shah et al., 2023). Herbold et al. run a series of linguistics counters to extract specific linguistic features. Meanwhile, Wu et al. explore numerous datasets to experiment with LLM-generated text detection, one of which is the TweepFake dataset (Herbold et al., 2023). The approach to detecting TweepFakes was by training numerous LLMs and classifiers implementing numerous methods on this dataset (Fagni et al., 2020). While the baseline detection in this study was relatively high, with accuracy scores ranging from 0.75 to 0.92 depending on the method and model, it would be interesting to compare counts of certain linguistic elements in the tweets, comparing the human Tweets to the bot-generated TweepFakes (Fagni et al., 2020). Herbold et al. took the approach of running numerous linguistic markers on a dataset of essays, some of which were human-witten and some of which were generated (Herbold et al., 2023). I chose to apply the linguistic marker analysis used by Herbold et al. on the TweepFakes dataset.

Table 1. Results

	Sentence complexity based on a certain number of dependency tags	Sentence complexity based on tree depth	MTLD lexical diversity score	Average number of epistemic markers per essay	Average number of nominalisations per essay
Bot	1.1799	4.4154	34.6195	0.01661	0.3296
Human	0.8387	3.4726	48.1929	0.01979	0.3388

II. Interpretation of Results and Limitations

The metric with greatest variability was the lexical diversity score. However, all of the other markers are relatively low per Tweet, which aligns with what one might expect from a short, not necessarily too complex bit of text.

It is important to consider this task and the results within the context of not only the metrics, but linguistic theory. For instance, tree depth is calculated algorithmically and left and right sides of trees are considered, which is not necessarily reflective of the universal nature of different languages, which may impact what linguistic trees look like in different disciplines (Yang, 2019). Following from this example and other properties, it may be a stretch to draw a direct parallel from counts of certain recognized patterns to linguistic theory. Additionally, it is interesting to note that the original TweepFakes paper utilized LLM methods to analyze the tweets, which is not necessarily explainable (Ray, 2023). While counts of linguistic patterns may not be a direct reflection of linguistic theories, it is directly explainable.

III. Future Directions & Linguistic Theory Considerations

There are numerous challenges that lie at the intersection and pose a disconnect in computational language application. Even in the context of Turing's imitation game, language has always been a notoriously intimidating challenge in AI (Turing, 1950; Bar-Hillel, 1953). **Human language** is "a dynamic, learned, hierarchical, relatively autonomous system of meaning-generating paradigmatic and syntagmatic signs that signify and communicate to self and others via speech communities and communities of practice throughout the life cycle" (Andrews, 2014 : 32). After all, language is how humans connect, constantly negotiate meanings, and establish cultural networks shaped by our experiences (Jakobson, 1987; Robbins, 2003). Bar-Hillel, a linguist, considers the trade-offs in machine translation (one of areas in natural language processing) between full automation and high quality.

Within the context of this specific task, considering the application of assessing computational language patterns in texts to determine whether they are generated may be more explainable than AI models, but both approaches do not reference the linguistics community of practice. "A community of practice is a group of people brought together by some mutual endeavor, some common enterprise in which they are engaged and to which they bring a shared repertoire of resources, including linguistic resources, and for which they are mutually accountable" (McConnell-Ginet, 2003 : 71). This is a continuous challenge in the language processing space, as it is an incredible challenge to bring together communities of practice when there exist discrepancies between priorities, funding, terminology, metrics, and best practices (Achler, 2023 : 3).

References

Achler, T. 2023. What AI, Neuroscience, and Cognitive Science Can Learn from Each Other: An Embedded Perspective. *Springer Nature.*

Andrews, E. 2014. Neuroscience and Multilingualism. *Cambridge, UK: Cambridge Universiy Press.*

Bar-Hillel, Y. 1953. Some Linguistic Problems Connected with Machine Translation. *Philosophy of Science*, 20, 217-225.

Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M. 2020. TweepFake: about Detecting Deepfake Tweets.

Herbold, S., Haulti-Janisz, A., Heuer, U., Kikteva, Z. & Trautsch, A. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13.

McConnell-Ginet, S. 2003. What's in a name: Social labeling and gender practices. In J. Holmes and M. Meyerhoff (Eds.), The Handbook of Language and Gender. *Blackwell.* 69–97.

Ray, P. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121-154.

Robbins, D. 2003. Vygotsky's and A.A. Leontiev's Semiotics and Psycholinguistics. *International Contributions in Psychology.*

Shah, A., Ranka, P., Dedhia, U., Prasad, S., Muni, S. et al. 2023. Detecting and Unmasking AI-Generated Content through Explainable Artificial Intelligence Using Stylistic Features. *West Yorkshire*, 14, 1043-1053.

Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind.* 49, 433-460.

Wu, J., Yang, S., Zhan, R., Yuan, Y., Wong, D. & Chao, L. 2024. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions.

Yang, J. 2019. Syntactic Hierarchy Depth: Distribution, Interrelation and Cross-Linguistic Properties. *Journal of Quantitative Linguistics; Abingdon*, 29, 129-145.