A Linguistic Evaluation of Machine-Generated "Real" and "Fake" News

Aziza Mirsaidova **Grace LeFevre**

CS 397 Seminar in Statistical Language Modeling
June 10, 2022

Abstract

Research on machine-generated fake news has often equated these two qualities, treating the task of identifying "machine-generated" news as equivalent to the task of identifying "fake news." In this project, we create datasets of machine-generated "real news" and machine-generated "fake news" by using GPT-Neo (Black et al., 2021) to perform text generation on input from the LIAR dataset (Wang, 2017). We have two goals: 1) to assess whether this approach is an effective way to create comparable machine-generated real and fake news, and 2) to ascertain if there are any detectable stylometric or linguistic differences between real and fake news generated in this way.

1 Introduction

The proliferation of fake news online has become a major concern in recent years. In particular, developments in neural language models have raised concern about machine-generated fake news being automatically disseminated on a large scale. For instance, Grover, a model for both generation and detection of fake news, successfully produced fake news rated more trustworthy than human-generated fake news (Zellers et al., 2019). More generally, the release of any powerful language models with the ability to produce human-like text, like GPT-3, comes with the potential risk of boosting the spread of fake news and disinformation (Floridi and Chiriatti, 2020).

Because of this risk, substantial attention has been paid to the linguistic features that might distinguish machine-generated news from human-generated news, but an underlying assumption of most work in this area has been equating "machine-generated news" with "fake news." Though a significant portion of machine-generated news is fake, automated text generation techniques have also been used regularly in authentic journalism throughout the past decade, for applications like producing stories from structured datasets (LeCompte, 2015). More recently, work has also been done to explore how AI techniques can contribute to even more difficult news-related tasks like investigative journalism (Stray, 2019). With this in mind, treating "machine-generated news detection" and "fake-news detection" as equivalent tasks is clearly an oversimplification.

This project seeks to fill a specific gap in this area: investigating stylometric or linguistic differences between machine-generated fake news and machine-generated real news. In one of the few projects that explored this question, Schuster et al. (2020) extended otherwise truthful human-generated news stories with machine-generated additions and analyzed whether stylometry could distinguish if the augmented stories were true or false. They concluded that, while stylometry can help distinguish between human-written and machine-written text, it was not useful for differentiating true and false machine-generated text. However, since the news Schuster et al. (2020) analyzed was mostly human-written, with machine-generated modifications or additions comprising only a small portion of the data, it remains unclear whether there are detectable stylistic differences between primarily machine-generated real news articles and primarily machine-generated fake news articles.

In this project, we explore this question by constructing a dataset of machine-generated real news and machine-generated fake news. We take sentences from the LIAR dataset (Wang, 2017) that were rated true and false as input to generate paragraphs of text using GPT-Neo. This process yields two sets of data, one machine-generated "fake news" data set and one machine-generated "real news" data set. We compare the two sets of articles using several linguistic features, including: named entities, referential words, and Zipf distributions.

Lastly, we use classification methods to predict whether the news articles we generated are real or fake news. We have two broad in this process. First, we want to assess whether our data generation strategy is a useful way to create datasets of machine-generated "real" and "fake" news. Second, we want to determine whether linguistic features can help in distinguishing real news articles and fake news articles that are both machine-generated by the same model.

2 Data Generation

2.1 The LIAR Dataset

LIAR (Wang, 2017) is a publicly available dataset for fake news detection. It consists of 12.8K human-labeled short statements occurring in various contexts collected from politifact.com. Based on the evaluation and justification of a politifact.com editor, each statement is labeled for truthfulness with one of six fine-grained ratings: pants on fire, false, barely true, half true, mostly true, and true. For the purposes of this project, we are only interested in the sentences labeled "false" and "true." Figure 1 shows an example of a false statement from the LIAR dataset.

Statement: *"Under the health care law, everybody will have lower rates, better quality care and better access."*
Speaker: Nancy Pelosi
Context: on 'Meet the Press'
Label: False
Justification: Even the study that Pelosi's staff cited as the source of that statement suggested that some people would pay more for health insurance. Analysis at the state level found the same thing. The general understanding of the word "everybody" is every person. The predictions dont back that up. We rule this statement False.

Figure 1: Example statement from LIAR data (Wang, 2017)

2.2 Text Generation Process

We performed text generation using GPT-Neo 1.3B, a pre-trained transformer model designed using EleutherAI's replication of the GPT-3 architecture (Black et al., 2021). GPT-Neo 1.3B was trained on the Pile dataset (Gao et al., 2020) for 380B tokens over 362,000 steps as a masked autoregressive language model using cross-entropy loss. This model is useful for our purposes because it was pretrained for generating texts from a prompt.

From the LIAR training dataset, we extracted the first 1000 true statements and the first 1000 false statements. Each true and false statement collected in this way was fed to this GPT-Neo model as a prompt for text generation. After experimenting with different parameter values, we settled on the following ones: max_length = 200, do_sample = True, temperature = 0.9, top_k = 50. This process yielded a paragraph of text for each statement, for a total of 1000 "fake news" paragraphs and 1000 "real news" paragraphs. An example of each is shown in Figure 2.

Label: False
Text: When Mitt Romney was governor of Massachusetts, we didn't just slow the rate of growth of our government, we actually cut it. Now, in our government spending, we still lag behind most of the world countries but it's not because of our political correctness. We simply have too much spending. Mitt was going to pay down the national debt. Yes, we were doing this on his watch and we did it by cutting our taxes and cutting benefits and cutting the size of our government. In our nation's history, not to pick a single thing, but we only did that for a while and at the end of the Obama administration the debt was reduced. But that's not the picture we have today. The picture we have today is that the federal government has grown, both the size of our government and the dollars we use to pay for it. In other words, with these two measures we have actually increased spending. Now, a lot ...

Label: True
Text: McCain opposed a requirement that the government buy American-made motorcycles. And he said all buy-American provisions were quote 'disgraceful.' The former senator then criticized Senator Ron Paul as being 'too anti-American.' However, when asked on March 1 about the possibility of the Senator supporting a bill to allow the U.S. military to use some of its funds to construct a new fleet of tanks, McCain said he 'would not rule that out.' On March 3, 2008, Senator McCain co-sponsored a bill that would have authorized President Bush to allow U.S. military forces to purchase up to ten advanced tanks (two of which were then under construction). The bill did not go anywhere before it was vetoed by Bush. In 2008, Senator McCain spoke at a meeting of the U.S. Conference of Catholic Bishops. While speaking on the subject of the global financial crisis, Senator McCain discussed the U.S., ...

Figure 2: Example "false" and "true" paragraphs from our generated data. The underlined sentence is the text that was provided as a prompt and the remainder was generated with GPT-Neo

2.3 Discussion

It's important to note a couple limitations of our data generation process. One limitation is that, since we specified a set sequence length, the paragraphs do not always end in a complete sentence. This is clearly not how paragraphs from human-generated real or fake news would be. Another important note is that, even though our "real" news examples were generated from statements rated "true" in the LIAR dataset, there is no guarantee that all of the information in the generated "true" paragraphs is actually true. For these and other reasons, we do not claim that our dataset consists of actual real and fake news but rather that it serves as a proxy for real and fake machine-generated news in the form of machine-generated text based on true or false input statements.

3 Linguistic Feature Analysis

We performed several types of linguistic feature analyses to compare our "real news" and "fake news" datasets, detailed below.

3.1 Named Entity Recognition

Named Entity Recognition (NER) is a widely used NLP technique that seeks to locate and classify named entities in a given text into pre-defined categories including person names, organizations, locations, quantities, expression of times, monetary values, etc. With the use of NER, we can extract main entities in a text which helps to sort unstructured data and detect important information from text. Prior to applying NER method to our dataset, we utilized NLTK's sentence tokenizer and removed special characters.

Spacy's Entity Recognizer pipeline assisted us in extracting entities from machine generated text. We found that there are 11,997 entities consisting of 16 unique entity labels in the false dataset, as shown in the upper bar chart in Figure 3. The total number of entities in the true dataset exceeded this, with 12,553 entities, as shown in the lower bar chart in Figure 3. In the fake news, ordinal, date, organization, nationalities and political groups, geopolitical and person entities are the most frequent entities with over one thousand occurrences in a sentence. The entities obtained from the true news differ somewhat from the false news with work of art, date, cardinal, person entities most frequent in the text.

These results show that entities related to both

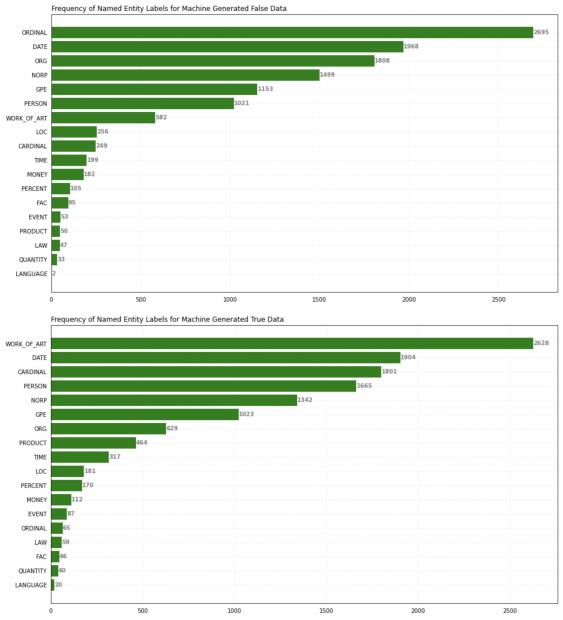

Figure 3: Named Entity Recognition distribution for unique occurrences of entities in "false" and "true" datasets.

true and false news differs in accordance with type of information involved in the dataset. For example, since false news commonly highlights the numerical and time information to bring more attention, it makes sense that the ordinal and date entities are the most frequent entities in the list. Further understanding the entities in both original and machine generated news may help with recognizing the differences as well as their contribution to fake news detection.

3.2 Referential words

Frequency of referential words in a text, like pronouns and proper nouns, has been examined in the context of fake news applications to see if it is correlated with truthfulness (e.g. Mahyoob et al., 2020). We used the Spacy package in Python to analyze referential term frequency by performing part of speech tagging and identifying pronouns and proper nouns. The results, as shown in Table 1, do not suggest any significant difference between the true and false data with respect to referential terms. The frequency of both pronouns and proper nouns is slightly lower in the true data, but likely not enough to be significant.

3.3 Zipf distributions

Zipf's Law states that, for a large enough corpus of text, the frequency with which a word is used in the corpus decreases with its rareness in the

	Pronouns	Proper Nouns
False data	0.04713	0.07648
True data	0.04392	0.07453

Table 1: Frequency of pronouns and proper nouns in "false" and "true" datasets, measured relative to all tokens present in each corpus

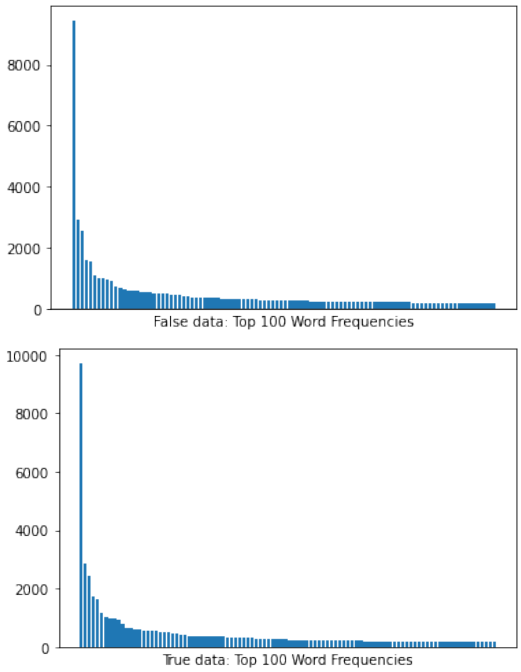

Figure 4: Zipf distributions for our "false" and "true" datasets, using the 100 most frequent words in each

text in an approximately hyperbolic way (Baayen, 2001). That is, the most frequent word in a corpus will appear twice as often as the second-most frequent word in the corpus, and so forth. Human language maintains this frequency distribution in an extremely reliable way cross-linguistically (Piantadosi, 2014). As a result, checking the Zipf distribution of a corpus can be a useful way of assessing how well it approximates human language.

Figure 4 shows the Zipf distributions for our generated corpora, using the 100 most frequent words in each. There don't seem to be significant differences between the two plots, but neither follow the frequency distribution that would be expected of natural language. For example, it is clear that the most frequent word is about three times as frequent as the second-most frequent word. This suggests that Zipf distributions remain an excellent way of detecting machine-generated text in general, but do not lend insight into distinguishing fake machine-generated text from true machine-generated text.

Model	**Accuracy Score**
Passive Aggressive Classifier	59.25%
Logistic Regression	58.19%
Naive Bayes	59.19%
Decision Tree Classifier	56.39%
BERT	50.13%

Figure 5: Classification models and their corresponding accuracy score detects the fake news based on machine generated "true" and "false" datasets.

4 Automatic Fake News Detection

We conducted fake news classification in our machine generated fake and true news dataset to understand performance of classification methods in machine generated text. Initially, we pre-processed the data creating a dataframe including false and true news in "text" column and binary labeled them in "labels" columns which shows whether text are representative of fake or true news. In addition, we counted the words using CountVectorizer and IDF and Tf-IDF which provides output in a sparse matrix representing the text. Then, we split the data into train and test sets training them in models shown in Figure 5.

As seen in Figure 5, Passive Aggressive Classifier exhibited the highest result in accordance with other classification techniques. It's an online learning algorithm that remains passive for correct classification outcome and aggressive to the lowest. Based on the accuracy score of the models, we can conclude that size of the dataset significantly effected the model performance to detect fake news.

5 Conclusion

This project highlights the importance of distinguishing the task of detecting "machine-generated" news from the task of detecting "fake news." In the future, we would like to extend our data generation process to create much larger machine-generated datasets and then perform more extensive linguistic feature analysis, with the goal of identifying features we could add to a discriminator model to better distinguish between machine-generated real and fake news.

References

R Harald Baayen. 2001. *Word frequency distributions*, volume 18. Springer Science & Business Media.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Celeste LeCompte. 2015. Automation in the newsroom.

Mohammad Mahyoob, Jeehaan Al-Garaady, and Musaad Alrahaili. 2020. Linguistic-based detection of fake news in social media. *International Journal of English Linguistics*, 11(1).

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Tal Schuster, Roei Schuster, Darsh J Shah, and Regina Barzilay. 2020. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510.

Jonathan Stray. 2019. Making artificial intelligence work for investigative journalism. *Digital Journalism*, 7(8):1076–1097.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.