Testing Computer Generated Disinformation with Human Written Fake News Using the Neural Model Grover

Aaron Chu-Carroll
Ardsley High School

Acknowledgements

I would like to thank Dr. Nasrin Mostafazadeh for offering her expertise and assistance. I would also like to thank Ms. Diana Evangelista and Ms. Jieun Yoo for their continued guidance and support. Finally, I would like to thank my parents for their constant help and motivation.

Table of Contents

List of Figures

Abstract

The “fake news epidemic” in recent years has presented a global challenge in modern society, misleading mass audiences and swaying public opinion. AI2's computer model Grover detects machine-written fake news by training a neural network with large amounts of data. This project aims to test how neurally-written news can be used to train a model for detection of all forms of fake news. Two rounds of testing occur with the same human-written real and fake news dataset, with different Grover models trained on different datasets. The first is trained with machine vs human written news data, while the second is trained with real vs fake human-written news data. Evaluating the performance of both models shows a statistically significant finding that neural news trained models cannot accurately detect human-written fake news, meaning that the characteristics of neural news are not the same as those of human news.

1 - Intr oduction

1.1 - Fake News

"Fake news" denotes any form of news intentionally designed to deceive. Although fake news has existed for as long as media news, it has become a recent problem as modern social media gives fake articles an opportunity to spread. Once a hoax or fabricated story is shared enough and becomes "viral", the damage is done, and no reporting or fact-checking can fix it. The top 1% of successful fake news reaches 1,000 to 100,000 people. The truth rarely diffuses to 1,000 (Vosoughi et al., 2019). This poses a larger issue than just small-scale deception of individual readers (Kossoff, 2018). Integrity of news is important for the conducting of business and politics. Fake news is the largest source of negative publicity towards businesses, such as the boycott of PepsiCo based on comments the CEO never actually made. The 2016 presidential debates were not focused on a central, agreed set of facts, but rather a debate of the falsity of information. Use of social media bots that appear as real users to perpetrate fake news further endangered this election's integrity (Bessi et al., 2016). The only method of preventing fake news is to identify it as fake before it spreads to a large audience. Although human fact-checking organizations such as Snopes and Politifact exist, they cannot keep up with the rapid output of fake news (Lazer et al., 2018).

1.2 - Natural Language Processing and Deep Learning

Natural Language Processing is the study of how computers can understand and interact with human language. The study of NLP has grown in popularity in the 2000s, coinciding with the rise of machine learning and deep learning neural models. A neural model is a set of algorithms that finds correlations . "Stacked" neural models, created of several layers, form a

deep learning model. Machine and deep learning focus on a computer system's ability to learn and process information. As a result, they are loosely modeled after the human brain, composed of a massive series of "nodes" (Lecun, 2015). A node is any location where a calculation or process occurs, comparable to human neurons. Parameters are any variables that are either given to the model or that the model learns during the training process. A deep learning model improves with the amount of layers, as it allows for more precise training, and therefore more accurate results. Since Natural Language Processing is a popular field, substantial research and creation has been recently published. OpenAI's GPT-2 is trained using language modeling, allowing it to write realistic human writing (Radford et al., 2019). Google's BERT is trained with masked language modeling, making it superior to GPT-2 for tasks such as Question Answering and Sentiment Analysis (Devlin, 2018). Question Answering is the automated answering of questions based on retrieved information. Sentiment Analysis is the analysis of how an author writes about a topic (anger, fear, enthusiasm). Grover is a deep learning model, loosely based on GPT-2, but specifically designed for the problem of fake news. Given a headline and news outlet, Grover can create a realistic article (Zellers et al., 2019). Grover can discriminate between its own neural fake news and human-written news with 92% accuracy. Unlike similar models, Grover has never been tested for its ability to discriminate between real and fake human news. Figure 1 shows Grover's ability to generate, as well as detect a machine-written fake news article.

Figure 1: Grover is capable of detecting and creating fake news

1.3 - Detection Strategies

Recent studies have taken various approaches to identify and suppress fake news. Finding unreliable articles at their source is more efficient and far easier than identifying per article (Baly et al., 2018). However, this strategy assumes that if a source publishes fabricated information once, they will consistently publish only fake articles. A dataset for Stance Detection — determining whether an author is in favor of or against a topic — to train a deep learning model has been tested, but this does not target real or fake news, instead finding positive or negative news. FakeDetector uses a traditional training deep learning approach (Zhang et at., 2018). The model is trained with a dataset of real and fake news, it identifies features and learns to discriminate between the real and fake data. Grover is trained the same way, but with a neural and human dataset. Grover can be trained and tested with any dataset that contains news articles and are annotated as real or fake.

1.4 - Datasets for Fake News

Creating a dataset of fake news is difficult, since it requires the creator to identify a large number of fake articles. FakeNewsNet uses articles collected from Politifact's lists of reliable and unreliable publications (Welch, 2017). McIntire created a dataset of fake news and real news

in a 1:1 ratio which is open for public access (McIntire, 2018). The fake data came from Kaggle's dataset of fake news, while the real data was scraped from various reliable news sources based on a list from AllSides. The dataset is ~6000 instances, annotated real or fake, with the source and authors.

2 - Statement of Purpose

The purpose of this study is to explore the role of AI in detection of neural and human fake news. The study determines whether a neurally-created news articles can be used to train a model that can effectively identify human-written fake news articles. First, Grover is trained with a dataset of real news articles written by humans and fake news articles automatically produced using Grover's own generation capability. This trained model is tested on McIntire's fake news dataset. The model is re-trained with the McIntire dataset, and a cross validation split is used to test with the same dataset. Comparative evaluations are completed to compare the accuracy of the first test and the second, including a t-test.

3- Methodology

3.1 - Initial Testing

Grover's discrimination mode requires training with a TPU (Tensor Processing Unit), a GPU works but not efficiently or reliably. Google's Colaboratory (Colab) application offers free use of a TPU without creating a virtual machine. All training occurs in Colaboratory on the TPU runtime setting. The first round of testing is performed with Grover's train data and Grover's test data. Although this result will not be used, it provides a simple way to test the environment and

resolve any problems before introducing outside data. Google's Cloud Storage program is used to store the train and test files, including the trained models. Grover's "large" size model is used for all trials and tests. Table 1 shows the difference in layers and parameters for model training between Grover's three sizes.

Grover Size	Layers	Parameters	Comparable to
Base	12	124 million	GPT and BERT-base
Large	24	355 million	BERT-large
Mega	48	1.5 billion	GPT-2

Table 1: Architecture parameters for the 3 model sizes

3.2 - Dataset Compatibility

The format of Grover's input data is not the same as the format of McIntire's dataset. Grover has many unnecessary annotations which are tagged null. A transformer algorithm is created in python, which modifies the .csv file to add annotation fields and reformat. The algorithm uses a for loop to change each line of the file. Lastly, it converts the file to a .jsonl, the same input file type as Grover.

3.3 - Training with Neural Fake News (NFN)

Once Grover has successfully run in Colab with its own data, outside data can be introduced. The model is trained already from previous tests, so retraining is unnecessary. The model is tested with the reformatted McIntire jsonl file. The run_discrimination.py file needs to be modified for data analysis. A for loop index command is added at the end of the code, then the file is reimported to Google Colab. This causes the program to output the exact values that

the model created, rather than just a percentage of the correct values. This is useful for data analysis to compare with the retrained model.

3.4 - Training with Human Fake News (HFN)

As a comparison, another Grover model is retrained with the McIntire dataset of human news. The dataset is made up of only 6336 instances. A 50/50 train/test split would not provide enough training data, and a 75/25 split would not provide enough testing data. Cross validation is used to test on all 6336 items without compromising the data. Cross validation is a method of testing on a smaller dataset without having to use a single train/test split, making the dataset even smaller. In a 5 fold cross validation, the model is trained and tested 5 times. Each time, 1/5 of the data is removed for testing, and the rest is used to train. As long as the same data is never tested twice, cross validation allows testing on the entire dataset without compromising the results. Figure 2 shows a representation of cross validation training and testing. In this study, a 5 fold cross validation is used. The model is trained and tested 5 times, where each trial should have approximately equal results.

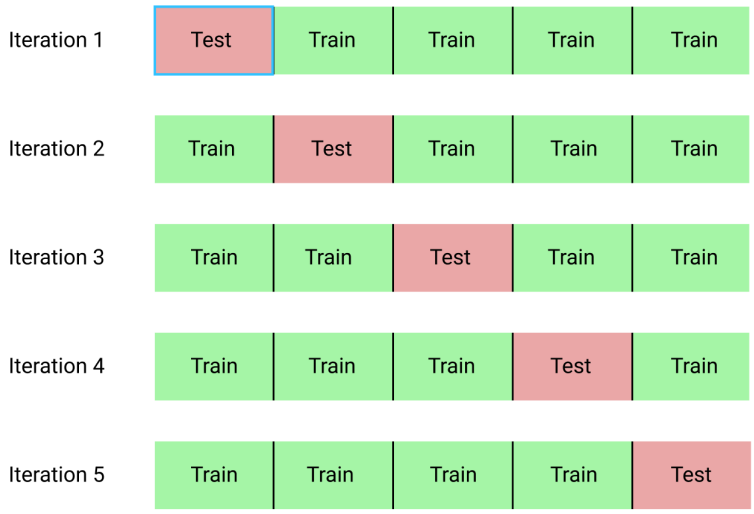

Figure 2: Cross validation tests one section at a time, training with the other data
https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85

3.5 - Data Analysis

The program in Colab returns a percentage and a list of values from an array. These values will be structured in the format “x :: y”, where each x and y value is a 1 (real) or a 0 (fake). X is the correct annotation, Y is the model’s prediction during testing. A t-test is performed with the values for the Neural Fake News (NFN) Test (Grover training, human testing), and the Human Fake News (HFN) Test (human training and testing).

4 - Results and Discussion

4.1 - Results

The neural fake news test scored 44.2% accuracy with the model trained by Grover’s input data. The human fake news test, trained and tested with cross validation on the same source data, scored a 95.4% accuracy. A t-test performed on the results from the two systems shows that the differences are significant at $p<0.001$. Table 2 shows the difference in training instances with the accuracy result of each test.

	NFN	HFN
# of Training Instances	10,000	6,336
# of Test Instances	6,336	6,336
Accuracy	44.2%	95.4%

Table 2: Experimental Results with dataset instances

4.2 - Results Interpretation

The HFN test was performed with the same source of train and test data (with cross validation so there is no train/test overlap in any fold), while the NFN test was performed with different train and test datasets. As a result, it is predictable that the HFN test would perform with higher accuracy than the NFN test. However, the HFN accuracy being over double that of the NFN accuracy — as well as the t-test results — show a statistically significant difference in the performance of the models. The 44.2% performance shows that the Grover training did not transfer to the human dataset testing. The characteristics of the neurally-created fake news must not be the same as those of the human-written fake news, otherwise the NFN model would have tested with higher accuracy. A neural model trained with neurally-created news cannot perform well with human fake news. In the future, neural fake news is very likely to become a large problem (this is why Grover was created in the first place). Since it has been determined that neural fake news has different features than human disinformation, the same trained algorithms will not be able to identify both types of fake news.

4.3 - Limitations of Study

The Grover model comes in three sizes: base, large, and mega. The study was originally meant to be performed using Grover's mega size, which has 48 layers and 1.5 billion parameters. Google Colaboratory's provided TPU runtime is limited in size and processing power. It was not able to run the mega size model, so the large model was used. The large model has 24 layers and 355 million parameters, which is still very large compared to other modern deep learning models for NLP.

4.4 - AI Challenges

This study is just one of many examples of a recent problem in the AI industry. As deep learning continues to grow, it has become practically required to use deep learning in some form to progress and tread grounds in the industry. Deep learning is far more taxing on computer hardware than machine learning or feature-based code. It typically requires the use of a GPU or TPU machine, and often requires the researcher to rent or purchase access to a larger GPU or TPU machine. This puts people in a lower economic class or with less available funds at a distinct disadvantage in AI research (Adams, 2019). It also allows large companies such as Google and IBM to dominate the AI industry, as researchers will often rely on them for funding. This can be harmful because it allows corporations to determine which research should be sponsored and which should not.

5 - Conclusion

5.1 - Overall Findings

The goal of this study was to determine how neural fake news training would perform with human fake news. This was to be performed using a state-of-the-art deep learning NLP model Grover. Two rounds of testing were performed, first with NFN training and second with HFN training, both were tested with the HFN dataset. NFN training resulted in a 44.2% accuracy, while HFN training results in a 95.4% accuracy.

Although neural models are capable of creating mass amounts of fake news, it has different characteristics than human-written fake news, and therefore cannot be used to train a model to detect human disinformation. This may present future issues or answer future problems when neural fake news makes its inevitable rise. Grover — and by extension, neural fake news

generation — clearly has a role in the future of fake news research. However, based on the results of this study, neural fake news is not a suitable replacement for human fake news, (at least in its current state), as it has entirely different characteristics per article.

5.2 - Future Research

The use of a second human dataset would have greatly benefited this study if it were not for the time and hardware limitations. Because of the nature of a train/test split, the HFN test was practically guaranteed to score higher accuracy than the NFN test. The study is still valid since the difference between the tests was so significant, but the HFN test being trained and tested on the same data source, while the NFN test was trained and tested on different data sources, presents an uncontrolled variable in the experimentation. A second dataset made up of human written fake and real news could remove this variable. While the current study proves that there is statistical significance, it is hard to say exactly how significant the difference is because of this variable.

Additionally, the Grover model has extreme potential for creation of datasets for training and testing of subsequent or unrelated models in the NLP field. Grover's generation mode was not used in this study, but it could be used to create massive datasets of neural fake news in relatively short periods of time. This is far more efficient than manually scraping websites from a list, which is how Grover was trained. This could allow other NLP models just as accurate as Grover (or more) to be created in a fraction of the time. Grover's neural fake news could be used in a study similar to this, but with a behavioral study, rather than computer science. This study found that neural fake news does not have the same features, or characteristics, as human fake news when tested with Grover. A similar study could be conducted but with volunteers to read

human fake news and Grover's fake news. This would determine whether neural fake news has different characteristics compared to human fake news when examined by an unsuspecting reader.

6 - Works Cited

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. “Language Models are Unsupervised Multitask Learners”. In: (2019).

[2] Adams, C. (2019, July 17). Expensive, Labour-Intensive, Time-Consuming: How Researchers Overcome Barriers in Machine Learning. Retrieved from https://journal.binarydistrict.com/expensive-labour-intensive-time-consuming-how-researcher-overcome-barriers-in-machine-learning/.

[3] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., & Nakov, P. (2018). Predicting Factuality of Reporting and Bias of News Media Sources. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

[4] Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday*, *21*(11). doi: 10.5210/fm.v21i11.7090

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language under-standing.arXiv preprint arXiv:1810.04805, 2018.

[6] Jiawei Zhang, Limeng Cui, Yanjie Fu, and Fisher B Gouza. 2018. Fake News Detection with Deep Diffusive NetworkModel.arXiv preprint arXiv:1805.08751 (2018)

[7] Kosoff, M. (2018, January 29). The Fake-News Epidemic Is Worse Than We Imagined. Retrieved from https://www.vanityfair.com/news/2018/01/the-fake-news-epidemic-is-worse-than-we-imagined.

[8] Lazer DMJ, Baum MA, Benkler Y et al.: The science of fake news.Science 359:1094–6, 2018x

[9] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015) doi:10.1038/nature14539

[10] McIntire, G. (2018, April 18). How to Build a "Fake News" Classification Model. Retrieved from https://opendatascience.com/how-to-build-a-fake-news-classification-model/.

[11] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News arXiv:1905.12616.

[12] S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. Science359, 1146–1151 (2018).

[13] Welch, R. (2017, April 20). PolitiFacts guide to fake news websites and what they ... Retrieved from https://www.politifact.com/punditfact/article/2017/apr/20/politifacts-guide-fake-news-websites-and-what-they/.