Data Generation for Neural Disinformation Detection

Tharindu Kumarage[1], Amrita Bhattacharjee[1], Kai Shu[2], and Huan Liu[1]

[1] Arizona State University, Tempe
{kskumara, abhatt43, huanliu}@asu.edu
[2] Illinois Institute of Technology, Chicago
kshu@iit.edu

Abstract. Incorporating large language models for various domain-specific NLP tasks has become prevalent due to the easy availability of pre-trained model checkpoints. However, fine-tuning these pre-trained models is necessary to improve performance on domain-specific tasks. Neural fake news detection is one such domain-specific task where the large language model needs to detect machine-generated fake news. Fine-tuning for a neural fake news detection task can be challenging since it requires collecting actual news articles and generating neural fake news counterparts. Therefore, in this paper, we explore the characteristics of the underlying data generation process of fine-tuning large language models for neural fake news detection. We present experiments to develop a deeper understanding of the fundamental properties of data generation. Some interesting findings have the potential to guide future research on neural fake news detection and to determine the quantity and variability of data required for fine-tuning large language models.

Keywords: Fake news detection · Data generation · Language models

1 Introduction

Large language models and the easy availability of publicly available pre-trained language models have driven a significant portion of the recent developments and advancements in the field of natural language processing. With the advent of open-source language models such as GPT-2 [14], Grover [16] etc, researchers are able to use much richer embeddings that incorporate contextual information, thus allowing better performances on many benchmark NLP tasks. However, these pre-trained models and embeddings may not be sufficient to tackle all the downstream NLP applications. This may be due to two different but related reasons: first, some domain specific tasks, such as analyzing medical scientific papers, would perform better if domain-specific embeddings are used over general embeddings such as these pre-trained models. Second, since these models were trained on older historical data, they tend to become out-dated over time [6]. This might be a significant problem for applications that need to learn about newly emerging concepts and topics (such as early detection of disinformation [11]). To test the efficacy of the pre-trained models on domain-specific

downstream tasks, and analyze the cost of fine-tuning required, we focus on *neural fake news detection* [16] as our downstream task. In this specific problem setting, we use these pre-trained models, train using our own data, and evaluate on a completely unseen dataset.

The neural fake news detection aims to read the body of a news article and determine whether it is human-written or machine-generated. Intuitively, better performance is expected when we employ more and more news articles to fine-tune the detector. However, data-generation is expensive and time-consuming. So in a real-world case, we might settle for a trade-off between detector performance and data generation costs. To explore and analyze the quantity and variability of data necessary for fine-tuning a natural language detection task, built on top of a pre-trained language model, we aim to answer the following research questions:

RQ1: How much new data do we need to generate in order to achieve reasonably good performance in the neural fake news detection?

RQ2: Given that we focus on the domain of neural fake news, how diverse do our news sources need to be in order to achieve a considerable performance gain?

In order to investigate and attempt to answer these research questions, we collect and generate necessary fine-tuning data (Section 2.1, 2.2) for our neural fake news detection framework (Section 2.3), design and perform experiments to analyze effects of dataset size and diversity (Section 3). We then talk about our experimental findings (Section 4) followed by some broader insights (Section 5), give a very brief overview of some related work (Section 6) and provide some discussion and future research directions in our concluding remarks (Section 7). The dataset and codes of our experiments are available at: https://bit.ly/3p9auhQ.

2 Fine-tuning for Neural Fake News Detection

We give an overview of the data collection, generation, fine-tuning and detection pipeline as shown in Figure 1 with details next.

2.1 Data Collection

We extract news articles on three topics: COVID-19, climate change and military ground vehicles. For each topic, we use a list of keywords to extract articles. Then we select 10 news sources for the article collection and scrape the relevant news articles from January 2020 to December 2020. A total of 14k articles are extracted and the source distribution of these articles is shown in Table 1.

2.2 Data Generation

The most important section of the dataset preparation effort in our analysis is to have the machine-generated data. Since we aim to optimize our performance on the evaluation data, we try to generate data 'similar' to the evaluation data. However, as mentioned previously, since we do not know the generation pipeline

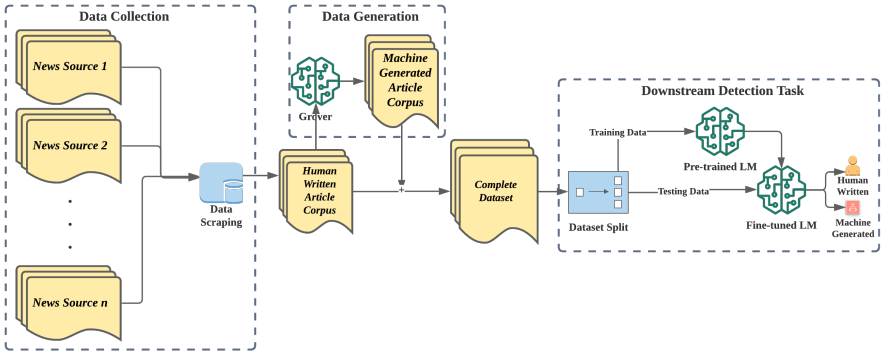


Fig. 1: Methodology Framework

Table 1: News sources used for data collection.

Source	Number of Articles
21st Century wire	257
Aljazeera	145
American Free Press	453
Canadian Dimension	169
Egypt Today	1049
France24	459
RT	476
Straits Times	5450
Taipei Times	2831
The Times of India	3112

for the evaluation data, at best we can estimate the steps and get a similar data distribution as the evaluation data. We exploited one of the state-of-the-art news article generators, Grover [16] as the generator. Using the Grover model [1] we generated 14k machine-generated articles, one for each of the 14k scraped news articles, under the following model configurations.

- **Conditioning** - We incorporated the conditioning provided by the Grover with the following attributes; headline, author(s), domain, and date.
- **Model size** - We used the *Grover Large* checkpoint for our generation.
- **Decoding method** - The decoding method controls how the model selects the next token while generating the text sequence. Here we used the top-p nucleus sampling (with $p = 0.96$) in Grover as the decoding method.

[1] https://github.com/rowanz/grover

2.3 Downstream Detection Task

Our downstream neural fake news detection task aims to determine whether an input news article is human-written or machine-generated. In this particular set of experiments, we use Grover as the generator, for our training data. To truly understand the effect of the fine-tuning dataset size and variation, we use a very simple detector framework. The detection module consists of a simple BERT-based encoder, where we initialize the weights using the pre-trained *bert-base-uncased* model. Then we have classification layers on top of this rich encoding that gives us the output probabilities.

3 Experiments

We perform experiments to answer the following research questions.

3.1 RQ1: Effect of Dataset Size

Here, our plan is to quantitatively analyze how much data should be generated in order to achieve considerably good performance in the downstream detection task. Therefore we created multiple sub-samples of our original dataset and used it to train the downstream task of neural fake news detection. The following proportions of the original data were used in the created sub-samples: 5%, 15%, 30%, 50%, 75%, and 100%.

3.2 RQ2.1: Effect of Number of News Sources

Under this experiment, we plan to quantitatively analyze how the diversity of our news sources affect the performance of the downstream detection task. We hypothesized that when the data samples within the training dataset are from different sources it would simulate the diversity effect. Therefore, we created multiple training sets that consist of news articles from different news sources. To have an increasing diversity we incorporated a top-k sampling approach where the news sources were ordered according to their original frequency within the full training dataset (Table (1)) and then created the multiple training sets by sampling from the top-k sources. We increased the value of k from 1 to 10 to include all the news sources in the original complete dataset. Moreover, to fix the effect of the size of the training dataset in the downstream task, we fix the sub-sample size. Each sub-sample dataset we created contains the same number of samples (size is defined according to previous experiment results; we use the 20% split of the original dataset).

Interestingly, as shown in Figure 2b, we do not see a correlation between the number of sources and the detection performance in this set of experiments. We hypothesize that this is due to the sources being very dissimilar, having dissimilar writing and reporting styles. This gives us hints that perhaps we should focus on these fine-grained lexical differences in the reporting style across various

sources, as opposed to simply varying the number of sources. To explore this hypothesis, we perform a thorough lexical analysis of the sources in the next set of experiments, as described in Section 3.3.

3.3 RQ2.2: Effect of Lexical Differences/Variations

Lexical Analysis of News Sources Here we analyzed the quality of the different news sources that we used to extract the news articles. The quality was measured in terms of *lexical readability* and *lexical richness*.

1. Readability - To measure the readability of the news articles, we employed three readability metrics; 1) Flesch Reading Ease, 2) Flesh-Kincaid Grade [5], and 3) Automated Readability Index (ARI) [10]. Flesch Reading Ease score is a value that could range from 0 and 100. The value bins are as below; post-college (0-30), college (31-50), high school (51-70), middle school (71-90), and grade 5 and/or below (91-100). Flesh-Kincaid Grade (FKG) is another similar measure of readability. However, the scale here is equivalent to the US grade level of education. This means the higher the score, the higher the quality. ARI score also yields a similar scale as Flesh-Kincaid Grade, yet the difference is that it incorporates characters per word to measure the reliability as opposed to the other two methods that use the syllables per word. These scores for each of the sources are noted in Table 2.
2. Richness - We measure the lexical richness of the articles extracted from the different sources in terms of lexical diversity. We used Type-token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD) [7] as the metrics for this analysis. TTR measures the diversity of the text in terms of average frequency of the unique words whereas MTLD measures the diversity as the mean length of sequential word strings in the text that maintains a given TTR. Since TTR is calculated as the mean of unique words used in the text, the value is susceptible to the increasing length of the articles. Therefore, we calculated the moving average TTR (MTTR) [1] so that it would be fair across all the sources. These scores for each of the sources are noted in Table 2.

Effect of Lexical Diversity on Detection Performance We again plan a similar experiment as Section 3.2 to quantitatively analyze how the lexical diversity of our news sources affect the performance of the downstream detection task. However, instead of increasing the number of sources, we incorporate the lexical features mentioned above to determine the source diversity. We employed a k-means cluster analysis on the 10 different sources in our dataset using the features extracted from the above lexical analysis. Each of the clusters from this experiment is considered as a bin of the same level of diversity in terms of text complexity and richness. Then the next step was to generate different training datasets from news sources within the same cluster to analyze the effect of the same level of diversity on the downstream task. Moreover, we created training

Table 2: News sources and corresponding lexical features.

Source	Lexical Readability			Lexical Richness	
	FRE	FKG	ARI	MTTR	MTLD
21st Century wire	21.05	20.70	23.46	0.9062	126.31
Aljazeera	18.46	35.51	43.30	0.9098	125.02
American Free Press	32.04	17.14	18.81	0.9066	128.39
Canadian Dimension	25.68	17.87	19.74	0.9002	114.61
Egypt Today	32.49	15.70	16.92	0.8832	88.59
France24	17.36	22.95	27.17	0.9118	135.82
RT	28.07	18.45	20.74	0.9111	138.28
Straits Times	10.375	26.74	32.26	0.9099	119.06
Taipei Times	42.01	14.04	15.12	0.9118	124.39
The Times of India	42.42	13.44	14.55	0.9055	111.00

datasets using news sources that reside in different clusters to analyze the effect of varying levels of diversity towards the downstream task.

3.4 Performance Evaluation

To evaluate the downstream detection task performance, we measure performance on two disjoint sets.

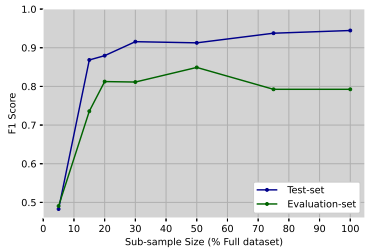
- Performance on the Test set - This is the hold-out set from our self-generated fine-tuning dataset
- Performance on the Evaluation set - This is a small dataset that consists of articles from the same source, yet the data collection and data generation pipelines are unknown to us. Therefore, performance on this dataset can be considered as the generalized performance of the detection mode.

4 Observations and Findings

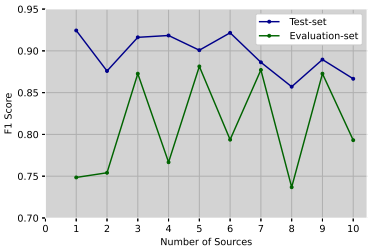
We provide some interesting observations and findings from the experiments.

Point of Diminishing Return Found. We observe that when training the downstream task on different sub-samples of the original dataset, there exists a point where accuracy starts to decrease. As seen in Figure 2a, the performance of the downstream detection task steadily increases with the dataset size, yet starts to display a decrease in gradient between 15% and 20% for both the evaluation and test datasets. This decrease confirms the existence of diminishing return in data generation for the neural fake news detection task.

Number of News Sources Doesn't Guarantee Increasing Data Diversity. We conducted two experiments to measure the effect of data source diversity on the neural fake news detection task. In our first experiment, we assumed



(a) Accuracy score variation of the downstream detection task when changing the training dataset size



(b) Accuracy score variation of the downstream detection task when changing the number of sources resides in the training dataset

Fig. 2: Experiments on RQ1 and RQ2

that the number of sources that reside in the dataset is positively correlated with the text diversity. Therefore when increasing the number of news sources in the training dataset we expected a linear trend in performance increase or decrease. However, according to the neural fake news detection performance as shown in 2b our assumption mustn't hold.

This observation opened up the following discussion on a more fine-grained analysis of the news sources to determine the dataset diversity.

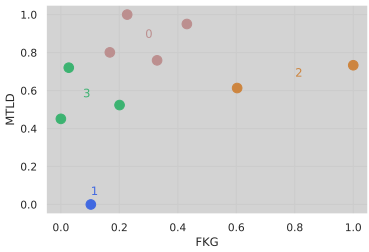
Linguistically Different Clusters of News Sources Found. As depicted in Figure 3a there exist 4 clusters of news sources within our dataset when parameterized using lexical readability and richness of the text.

From Table 3 we can see the news source assignment for each of these clusters and this observation highlights the necessity of a lexical analysis to determine the dataset diversity to answer the second research question of our work.

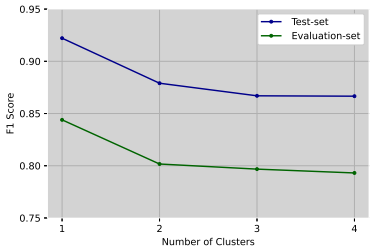
News Source Diversity Can Hinder the Detection Performance. After finding lexically different clusters of news sources, we conducted another experiment outline in section 3.3 to quantitatively analyze the effect of news source diversity on the neural fake news detection task. Each cluster of news sources can be considered as a different level of diversity and we evaluated the average

Table 3: Cluster assignment of news sources.

Cluster	News Source(s)
0	21st Century wire, france24, American Free Press, RT
1	Egypt Today
2	Aljazeera, Straits Times
3	Canadian Dimension, Taipei Times, The Times of India



(a) k-means clustering on the lexical features extracted for each of the news sources - X axis represent readability (Flesh-Kincaid Grade) and Y axis represent diversity (Measure of Textual Lexical Diversity).



(b) Accuracy score variation of the downstream detection task when changing the number of clusters resides in the training dataset.

Fig. 3: Lexical feature analysis experiments on news sources

detection performance when 1) training data is created by sampling articles only from the same cluster - the number of cluster equals 1 case in Figure 3b and, 2) training data is created by sampling articles from different clusters - the number of cluster equals 2-4 cases in Figure 3b. It's apparent that when the diversity increases the performance of the neural fake news detection decreases.

5 Discussion

Achieving Adequate Performance Under Limited Resources for Data Generation. In many machine learning tasks, there is often a trade-off between the cost of resources for data generation and the model performance. From the observations on diminishing return that we obtained in section 4, it is evident that there exists a sweet spot where we could gain good enough performance for neural fake news detection with a comparably smaller amount of data, given we built our detector on a pre-trained large language model. Therefore, we could presume that even with limited resources for data generation, achieving satisfactory performance in the neural fake news detection task is feasible.

Meticulously Planned Data Collection Process and Preliminary Analysis Could Further Reduce the Cost of Data Generation. From the observations we obtained under experiments on RQ2, it is determined that blindly increasing the number of news sources in detection may not result in a performance gain, unlike our expectation. A thorough lexical analysis on data quality would be the key as we have seen in the section 4. The quality and fine-grained variability of news sources is an unquestionably important factor as opposed to quantity. Therefore, conducting a preliminary analysis on the potential news

sources could reduce the overhead of generating news articles on a large number of news sources that might not add value to the latter task of detecting the neural fake news.

6 Related Work

Detection of deep-text, i.e. machine-generated text, has become one of the emerging areas of research mainly due to the recent advancements in pre-trained large language models [2]. Numerous pre-trained large language models such as Grover [16], GPT-2 [14], GPT-3 [8], T5 [9], and CTRL [4] came into the arena of language modeling within the last couple of years and overthrown the previous state-of-the-art performances in natural language generation. Consequently, there exists a potential danger of malicious use of these models in producing text that could lead to harmful circumstances such as deep headline [13], deep tweets [2], and fake news generation (i.e. neural fake-news) that has the potential to drive disinformation campaigns [16].

Generators such as BERT, RoBERTa, GPT-2, Grover [3, 14, 16] have shown good performance in detecting machine-generated text in many domains. One common requirement for the aforementioned detection efforts is the dataset preparation for training the detector model. For the tasks like fake news detection, there has been comprehensive dataset preparation [12] as opposed to neural fake news detection. Particularly, generating the machine text is a challenging task. There exist some public datasets creation efforts such as neural news [15], Grover generations of real-news [16], and GPT-2 generations of Webtext [14]. The first two datasets consist of machine-generated news articles by the Grover model and the last one consists of machine-generated text from GPT-2. Our work carefully follows these dataset preparation efforts while analyzing the fine-grained characteristics of the dataset requirement and provides observations for future dataset preparations demanded by neural fake news detection.

7 Conclusion

We studied the process of data generation to support fine-tuning of large language models to detect neural fake news detection, i.e. machine-generated fake news. We found that the neural fake news detection model fine-tuning can be supported even under the low resources for data generation. Furthermore, we established the importance of lexically analyzing the data sources ahead of the data generation process that would help mitigate unwanted use of data generation resources. We believe our work paves the way for a better understanding of the specific requirements of the dataset generation process to aid neural fake news detection in future. One natural extension of our work is to analyze the effect of adding more generators into the dataset and how that would affect the fine-tuned performance of the neural fake news detectors.

Acknowledgements This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001120C0123. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

References

1. Covington, M.A., McFall, J.D.: Cutting the gordian knot: The moving-average type–token ratio (mattr). Journal of quantitative linguistics **17**(2), 94–100 (2010)
2. Fagni, T., Falchi, F., Gambini, M., Martella, A., Tesconi, M.: Tweepfake: About detecting deepfake tweets. Plos one **16**(5), e0251415 (2021)
3. Ippolito, D., Duckworth, D., Callison-Burch, C., Eck, D.: Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650 (2019)
4. Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C., Socher, R.: Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858 (2019)
5. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
6. Lazaridou, A., Kuncoro, A., Gribovskaya, E., Agrawal, D., Liska, A., Terzi, T., Gimenez, M., d'Autume, C.d.M., Ruder, S., Yogatama, D., et al.: Pitfalls of static language modelling. arXiv preprint arXiv:2102.01951 (2021)
7. McCarthy, P.M., Jarvis, S.: Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods **42**(2), 381–392 (2010)
8. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
9. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
10. Senter, R., Smith, E.A.: Automated readability index. Tech. rep., CINCINNATI UNIV OH (1967)
11. Shu, K., Awadallah, A.H., Dumais, S., Liu, H.: Detecting fake news with weak social supervision. IEEE Intelligent Systems (2020)
12. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big Data **8**(3), 171–188 (2020)
13. Shu, K., Wang, S., Le, T., Lee, D., Liu, H.: Deep headline generation for clickbait detection. In: 2018 IEEE International Conference on Data Mining (ICDM). pp. 467–476. IEEE (2018)
14. Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., et al.: Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203 (2019)
15. Tan, R., Plummer, B.A., Saenko, K.: Detecting cross-modal inconsistency to defend against neural fake news. arXiv preprint arXiv:2009.07698 (2020)
16. Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., Choi, Y.: Defending against neural fake news. arXiv preprint arXiv:1905.12616 (2019)