Real vs Fake : Guess we might know!
Pretrained Electra for Human vs Machine Generated Text

Ana Cheyre **Anery Patel** **Ayushi Mishra** **Devyani Gauri**

Abstract

Machine generated text, despite having received a lot of popularity in recent years, has not yet reached the quality and naturalness of human writing. With this project we aim to explore the task of identifying the difference between human versus machine-generated text using ELECTRA (an encoder-based model) over various language characteristics and compare that with the existing baseline models available for this task. Pertaining to the domain of fake news detection, we observe the computational advantage ELECTRA has over our baseline model and find its accuracy to be comparable to the .

1 Introduction

With the ubiquity of information in the internet era, it has become more likely for a reader to come across fake news and articles on a daily basis. Can social media be a reliable news source? It has become an almost inevitable element of our culture. Not all of the time. As quickly as the news can be accessed, fakenews spreads at a comparable speed, infact at times faster than the real news (blo).

To contribute to the endeavors, we address this issue by proposing to leverage a pretrained Electra model(Clark et al., 2020) on the WELFake News dataset (Verma et al., 2021). The original paper presents a new training task for Electra in which the model learns to distinguish real input tokens from plausible but synthetically generated replacements. This gives the model its power to act as a discriminator. Since our focus is on a binary label classification task to predict real or fake news article, Electra is a reasonable choice when it comes to using a language model for text classification. We evaluate Electra on WELFake dataset and further compare it with the predictions generated by BERT.

2 Related Work

A lot of text generation models already exist and have been used for a variety of tasks. Some existing models are GPT-2 (Radford et al., 2019), GROVER (Zellers et al., 2019), GPT-3 (Brown et al., 2020).

BERT (Devlin et al., 2018) has been used as a detector for machine-generated text. On the WELFake Dataset, BERT has an accuracy of 93.79% (Zellers et al., 2019). However, we computed our results with the BERT classifier for this task.

There is also work that has been done to identify text generated by machines (or fake text), with state-of-the-art results by the GROVER model on the RealNews dataset (Zellers et al., 2019). Since RealNews is a huge dataset with a large computational complexity associated with it, we use the WELFake Dataset (Verma et al., 2021) for our task. The WELFake model in this work was the best performing model on the dataset so far.

3 Model

We decided to use ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) for the task of classifying a text between machine-generated and human-generated.

ELECTRA uses replaced token detection

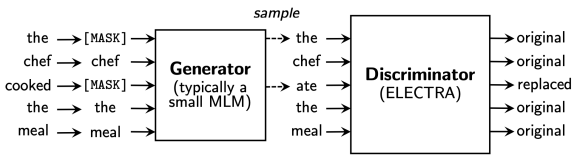

Figure 1: ELECTRA model architecture

that trains a masked language model (MLM) while learning from all input positions (like a Language Model). Similar to generative adversarial networks (GANs), ELECTRA trains the model to distinguish between "real" and "fake" input data generated using the generator (typically a small MLM). The model learns from all input tokens instead of just the small masked out subset, making it more computationally efficient.

There are three released models at this time: ELECTRA-Small (12 layers, 256 hidden layers) ELECTRA-Base (12 layers, 768 hidden layers) ELECTRA-Large (24 layers, 1024 hidden layers)

To improve the efficiency of pre-training, (Clark et al., 2020) proposed developing a small model that can be quickly trained on a single GPU. According to their results, it is more efficient and obtains better results than the BERT model of the same size.

We decided to use ELECTRA-Small for our task. We set the learning rate to 0.0001, used an Adam optimizer for gradient descent, maximum embedding length of 512, and kept the batch size at 8 for 5 epochs. The objective function is determined by the CrossEntropy loss used for binary classification.

4 Dataset

The dataset used is WELFake open dataset, which incorporates four popular datasets (Kaggle, McIntire, Reuters, BuzzFeed Political) to generate an unbiased classification output. It contains about 72,000 articles with 35,028 real and 37,106 fake news articles. As seen in Table 1, WelFake consists of articles primarily from Reuters and Kaggle, some articles from McIntire and the least number of articles from BuzzFeed Political. Each of the constituent datasets is balanced on their own, hence making WelFake a balanced dataset as a result.

WELFAKE DATA SET

Dataset	*Real news*	*Fake news*
Kaggle	10387	10413
McIntire	3171	3164
Reuters	21417	23481
BuzzFeed Political	53	48
WELFake dataset	**35,028**	**37,106**

Figure 2: WELFake dataset composition as given in the original paper

From Figure 2, we can notice the following: Short sentences are more prevalent in real news than in fake news. Readability index of Real news is slightly higher than that of Fake news. Fake news articles have more subjectivity than real news articles. Number of articles is higher for real news than fake news.

The dataset contains four columns: Serial number (starting from 0) Title (news heading) Text (news content) Label (0 = real and 1 = fake)

5 Experiments

We finetune and evaluate Electra-small and BERT (Devlin et al., 2018) on WELFake dataset for the machine generated versus human generated text classification task. The 'label' column of the WELFake dataset is set as the target variable with '1' denoting fake text and '0' denoting a real text.
We compare the results of Electra with BERT. The model proposed by WELFake data set achieves a fake news classification accuracy of up to 96.73% (Verma et al., 2021). We implemented our models on a Google Colab notebook that uses a single 12GB NVIDIA Tesla K80 GPU. We experiment with each model on a randomly shuffled dataset of train-test combinations of 70%-30% and 80%-20%. The Electra model is validated on different sets of hyperparameters namely learning rate, number of epochs, max_length of the sequence and batch size to identify the optimal setting for efficient outcome.

FAKE VERSUS REAL NEWS DISTRIBUTION IN WELFAKE

Category	*Real news [%]*	*Fake news [%]*
Short sentences	60.9	39.1
Readability index	51.7	48.3
Subjectivity	45.4	54.6
Number of articles	53.9	46.1
WELFake dataset	**48.55%**	**51.45%**

Figure 3: WELFake text distribution as given in the original paper

The objective function is determined by the CrossEntropy loss used for binary classification. To provide a fair comparison, we also train a BERT-Small model using the same hyperparameters. To evaluate the models, we implement accuracy, precision, recall and F1-score on the test set for each experimental setting with hyperparameter variations.

6 Results

6.1 BERT baseline results

The results for our baseline model (BERT) along with its various experimental setups can be seen in Figure 4. BERT-Standard has better accuracy and F1 scores than all other hyperparameter settings except Batch Size 8 and Training Size 80%. We observe that increasing the learning rate to 1.00E-4 has had quite an impact on our scores, decreasing by 51.3% in terms of accuracy for the BERT model. The F1 score and Recall scores decreased drastically by 75.25% and 83.53% respectively. Significant differences aren't observed in accuracy when comparing the BERT-Standard with other hyperparameter changes. The training time for the BERT-Standard model was around 5 hours. And for some hyperparameter cases, it exceeded up to 6 hours with P100 GPU and High RAM settings.

6.2 Electra model results

The results for our ELECTRA model along with its various experimental setups can be seen in Figure 5. The ELECTRA-Standard has better accuracy and F1 scores than all other hyperparameter settings except learning rate 2.00E-5. We observe that reducing Epochs

	BERT			
	Accuracy	F1	Precision	Recall
BERT-Standard	0.994	0.994	0.992	0.996
LR 1.00E-5	0.994	0.994	0.995	0.993
LR 1.00E-4	0.484	0.246	0.490	0.164
Max length 256	0.993	0.993	0.996	0.991
Batch size 8	0.995	0.995	0.995	0.995
Epochs 3	0.992	0.992	0.996	0.988
Training size 80%	0.995	0.995	0.996	0.994

Figure 4: BERT model ablations over the test set of WELFake Dataset

to 3 and Max length to 256 reduces the accuracy and F1 score to some significant levels. The training time for the ELECTRA-Standard model was around 1 hour and 40 minutes with P100GPU and High RAM settings.

	ELECTRA			
	Accuracy	F1	Precision	Recall
ELECTRA-Standard	0.991	0.991	0.996	0.989
LR 2.00E-5	0.992	0.991	0.990	0.995
Max length 256	0.987	0.985	0.984	0.990
Batch size 16	0.993	0.993	0.994	0.993
Epochs 3	0.985	0.983	0.983	0.988
Training size 80%	0.990	0.988	0.989	0.991

Figure 5: ELECTRA model ablations over the test set of WELFake Dataset

6.3 Comparing baseline with model results

Even though the BERT-Standard model's accuracy is higher than that of the ELECTRA-Standard model by 0.003, the computational advantage of ELECTRA over BERT is greater. Thereby, making it a comparable model with higher computational efficiency. It should be noted that to compensate for any overfitting issue, we have tried different split ratios (70-30 and 80-20) for training the dataset when hyperparameter tuning did not show much difference in results and tried random shuffling for ELECTRA. We assume that the dataset itself might have an inherent bias that could be a potential reason for us observing high-performance scores.

7 Conclusion

In this work, we showed that ELECTRA can be a good alternative to BERT for human vs machine-generated text tasks, especially in terms of computational efficiency. Given the results, it would be good to be able to apply the same models to another machine-generated vs human database in future work. One can even create their database to ensure that there is no bias introduced into the dataset. Using different model releases of ELECTRA (base and large) to see whether they outperform BERT and WELFake models would be another avenue worth investigating.

References

Mit sloan research about social media, misinformation, and elections. https://mitsloan.mit.edu/ideas-made-to-matter/mit-sloan-research-about-social-media-misinformation-and-elections.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. 2021. Welfake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893.

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.