Misinformation Classification Using BERT models

Kris Riedman krisr@berkeley.edu

December 2023

Abstract

Larger models produce better results but not everyone has the resources necessary to run those models. I conducted a series of text classification experiments testing different sized BERT models on news articles that may contain misinformation. I found that smaller BERT models offer similar results as compared to larger BERT models.

Introduction

Misinformation in the news is a real problem for our society. A seemingly legitimate news outlet can pass falsehood as fact. In addition to this, the tools that can help address this problem are out of reach for many. Misinformation detection is important for several reasons. Recent elections have shown that bad actors will spread false information to help their candidates win office. We have also seen that misinformation can affect public health with regards to vaccinations and the spread of communicable diseases such as measles and Covid-19. Finally, with the invention of large language models, it is easier than ever to create vast amounts of computer generated text that can be tailored to spread any number of misleading claims that masquerade as fact.

Currently, there are a plethora of large models that can determine if an article contains misleading information but these models are either very large or return a lot of extra information that can be useful in some cases (model explainability) but certainly complicate the overall message. My goal is to identify a smaller model that will run quickly and efficiently even in a constrained resource environment. Furthermore, the results will be simple and easy to interpret: this article does or does not contain misinformation.

Background

PANACEA is an online tool created by a group of researchers out of Great Britain that “provides automated veracity assessment and supporting evidence for the input claim.” PANACEA has a 94% macro F1 score for veracity classification which is higher than the GEAR model used for the same task. This model may be useful for users who wish to understand why a claim is spurious and the facts surrounding it, but for others this may amount to information overload.

Another common approach to propaganda detection is sentence level and fragment level propaganda detection as described in “Neural Architectures for Fine-Grained Propaganda Detection in News” by Gupta et al. Their model was ranked 3rd in the fine-grained propaganda detection shared task 2019 specifically for the FLC task with an F1 score of 19%. Unfortunately, this model is geared towards sentence and fragment level analysis and may not scale to entire news articles.

A different team identified eighteen different propaganda techniques in their paper titled “Fine-Grained Analysis of Propaganda in News Articles". Their model attained a F1 score of 22.58% for the combined task of identifying propaganda spans and assigning the correct label. Identifying spans of propaganda

embedded in text can help with model explainability, but the nuance between different propaganda types may be too fine grained for most users.

In their paper "Proppy: Organizing the News Based on Their Propagandistic Content", Barron-Cedeno et.al. used a maximum entropy text classifier in a few experiments including a 4-way classifier: trusted vs. propaganda vs. hoax vs. satire, and a 2-way classifier: propaganda vs. non-propaganda. This model attained an average F1 score of 66.60% when predicting if an article contains propaganda. A potential improvement on this model would be to utilize a transformer architecture which is what I have set out to do.

Methods

The primary goal of this paper is to determine if smaller BERT models are as effective as the BERT base model. The secondary goal is to reduce certain parameters such as batch size and number of tokens per sample to see if the smaller models are still performing optimally. My data is sourced from multiple news outlets that have been labeled as 'true' and misinformation. The models I have selected do not include any veracity assessment module so we must instead rely on analyzing the language and tone of each article.

An example of a 'true' article that does not contain misinformation:

"The head of a conservative Republican faction in the U.S. Congress, who voted this month for a huge expansion of the national debt to pay for tax cuts, called himself a "fiscal conservative" on Sunday and urged budget restraint in 2018. In keeping with a sharp pivot underway among Republicans, U.S. Representative Mark Meadows, speaking on CBS' "Face the Nation," drew a hard line on federal spending, which lawmakers are bracing to do battle over in January."

An example of an article that contains misinformation:

"Donald Trump just couldn't wish all Americans a Happy New Year and leave it at that. Instead, he had to give a shout out to his enemies, haters and the very dishonest fake news media. The former reality show star had just one job to do and he couldn't do it. As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year, President Angry Pants tweeted. 2018 will be a great year for America! As our Country rapidly grows stronger and smarter, I want to wish all of my friends, supporters, enemies, haters, and even the very dishonest Fake News Media, a Happy and Healthy New Year. 2018 will be a great year for America! Donald J. Trump"

A large BERT model is able to classify each example with a high degree of accuracy but will require a considerable amount of compute resources to train. The BERT base cased model I am using is around 400 mb in size and takes around 45 minutes to fine tune on 10k training examples. A smaller model with a similar accuracy will not only be fine tuned faster but will also take less resources ensuring a wider audience will have the ability to use this product.

I will perform a comparative analysis of different sized models by fine tuning and evaluating three different models with three different sets of hyperparameters. The three models I decided to test are BERT base cased, DistilRoberta, and DistillBert. The training set is half with the true label and half with the false label. I created a model for each combination of two hyperparameters: max length 200 and 500, batch size 4 and batch 8. To ensure a fair assessment during experimentation, each model was trained on the same preprocessed data. To evaluate the effectiveness of the model, I predicted the label on new articles that were never seen. The evaluation data includes a third group of articles sourced from Russian national news outlets which are all labeled as

misinformation. By comparing the results from each model and experiment, I can determine if the smaller models are as capable as a larger model at producing quality results. I will measure success by comparing the F1 scores from each model. I will consider this experiment a success if the smaller models' F1 score are within five percentage points of the best baseline model F1 score.

	200 Max length		500 Max length	
	Batch Size 4	Batch Size 8	Batch Size 4	Batch Size 8
Bert	0.53	0.17	0.58	**0.99**
DistilRoberta	**0.97**	0.17	0.84	1.00
DistilBert	0.48	0.29	**0.94**	**0.97**

Results and discussion

The baseline model I used was a BERT base cased transformer with a batch size of 8 and a max length of 500. All three models using the same hyperparameters as the baseline model were able to achieve excellent performance when predicting the label on the evaluation set with an average F1 score of ~98%. The fact that both of the smaller models obtained similar F1 scores to the baseline model is a huge success.

DistilRoberta did very well with max length 200 and batch size 4 compared to both the baseline model and the DistilBert model. This improvement could be attributed to the extended and expanded pre-training on DistilRoberta compared to other Bert models.

DistilBert was the best model with max length 500 and batch size 4. DistilRoberta also performed well with those hyperparameters. Since both of these models are more efficient, they are better able to take advantage of the smaller batch size.

Evidently there is an issue involving a max length of 200 and a batch size of 8 since all models had very low F1 scores. A possible explanation for this could be that a larger batch size is causing the model to overfit the training data or it could be an issue with the model getting stuck in a local minima during gradient descent.

DistilBert and the baseline model benefit from a larger max length size. This makes sense since DistilBert has a very similar architecture to the Bert base model. A larger max length size allows the models to more accurately learn from each training example.

All models with the max length 500 and batch size 8 failed to properly classify these examples. The most likely explanation has to do with the sample length: it is far too short for the model to accurately classify. Here are two examples that were misclassified by all three models:

"The migration wave is part of preparations for the New World Order, says analyst David Icke."

"Polish policy towards Russia is entirely based on the support of the United States."

Here is another misclassified example that indicates a data issue: this training example shouldn't have been included in the dataset since it really isn't news but it also includes some extra verbiage that doesn't correspond to language but instead are probably artifacts from when the dataset was collected.

"Next Swipe left/right Koala gets so excited it runs head-first into a tree Mason, a baby Koala at the Port Stephens Koala sanctuary in Australia, gets so excited he runs head-first into a tree."

Another data issue this time involving an untranslated article. Obviously a language model trained in English would have difficulty in correctly classifying an article in a different language.

“españa , almirante kuznetsov , ceuta , portaaviones Imagen del portaaviones Almirante Kuznetsov. Fuente:Reuters
La ruta y la misión del portaaviones Almirante Kuznetsov, según el secretario de prensa del presidente de la Federación de Rusia, Dmitri Peskov, se encuentran en un sobre cerrado en el que se lee “Completamente secreto”.
Sin embargo, según informaba anteriormente el Ministerio de Asuntos Exteriores español, se preveía que el portaaviones llegara al puerto de Ceuta el 28 de octubre . En Madrid señalan que el permiso correspondiente se expidió el pasado mes de septiembre.
No obstante, en cuanto la noticia sobre la inminente llegada de los buques a Ceuta llegó a la prensa, España comenzó a recibir críticas de sus aliados en la OTAN.”

This misclassification highlights perhaps the largest issue with this dataset: the label of truth and misinformation is based on news outlet as opposed to article content. This inevitably does introduce some bias into the model; however, it is a fact that articles labeled as misinformation are authored by news outlets that have a higher probability of containing misinformation compared to other news outlets and vice versa. This is an example of an article written by an otherwise reputable news outlet that contains misinformation in the form of a quote.

“The Turkish government s spokesman on Wednesday said that the United States decision to recognize Jerusalem as the capital of Israel will plunge the region and the world into a fire with no end in sight . Declaring Jerusalem a capital is disregarding history and the truths in the region, it is a big injustice/cruelty, shortsightedness, foolishness/madness, it is plunging the region and the world into a fire with no end in sight, Deputy Prime Minister Bekir Bozdag said on Twitter. I call on everyone to act logically, respect the agreements they signed and behave reasonably, avoid risking world peace for domestic politics or other reasons, he said. U.S. officials have said President Donald Trump is likely to give a speech on Wednesday unilaterally recognizing Jerusalem as Israel s capital, a step that would break with decades of U.S. policy.”

The greatest result of this experiment is that the DistilBert and DistilRoberta models performed as well and in some cases better compared to the baseline model which proves the hypothesis that a smaller model can return solid results.

Conclusion

Classifying news articles as misinformation efficiently is difficult. The Distilbert model and the DistilRoberta model can reliably identify misinformation using a fraction of the training time and the models are up to 40% smaller than the BERT base model. Further optimizations can be made to the Distilbert and DistilRoberta models through prefix training using Lora. This will further decrease the overall size of the models. Additionally, I would spend more time on experimenting and optimizing hyperparameters for each model. Finally, this dataset could be improved by removing foreign language articles, tweets, and artifacts from the data collection process.

References

Pankaj Gupta, Khushbu Saxena, Usama Yaseen, Thomas Runkler, and Hinrich Schütze. 2019. Neural Architectures for Fine-Grained Propaganda Detection in News. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 92–97, Hong Kong, China. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav

Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Runcong Zhao, Miguel Arana-catania, Lixing Zhu, Elena Kochkina, Lin Gui, Arkaitz Zubiaga, Rob Procter, Maria Liakata, and Yulan He. 2023. PANACEA: An Automated Misinformation Detection System on COVID-19. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 67–74, Dubrovnik, Croatia. Association for Computational Linguistics.

Alberto Barrón-Cedeño, Israa Jaradat, Giovanni Da San Martino, Preslav Nakov. 2019. Proppy: Organizing the News Based on Their Propagandistic Content. In *Information Processing & Management, Volume 56, Issue 5,* Pages 1849-1864, United States. Pergamon Press, Inc.

Appendix

- Dataset - Misinformation & Fake News text dataset 79k (kaggle.com)
- BERT models - google-research/bert: TensorFlow code and pre-trained models for BERT (github.com)
- Hugging face BERT: BERT (huggingface.co)
- Hugging face DistilRoberta: distilroberta-base · Hugging Face
- Hugging face Distilbert: DistilBERT (huggingface.co)
- Code: riedmank/w266_final_project (github.com)