Conditional Fake news generation
Machine Learning for Natural Language Processing 2021

Matthieu Futeral-Peter
ENSAE Paris
matthieu.futeral.peter@ensae.fr

Alexandre Rio
ENSAE Paris
alexandre.rio@ensae.fr

Abstract

Recently, neural auto-regressive language models, and particularly transformer-based models, have shown promising results in generating plausible text. However, despite the advances, it is hard to control the generated sequence. Indeed, to generate text, the user has to give a prompt to the model that generates text based on that prompt. Except that, the user has no control over the generated sequence. Some models try to handle that problem by conditioning the language model (Keskar et al., 2019). In this project, we exploit that idea in order to generate fake news conditionally to attributes: a category (business, technology, health or entertainment), some keywords and a title. The code and the notebook of this project can be found here[1].

1 Problem Framing

Our objective is to build a conditional neural language model to generate fake news and make sure our model does generate sequences according to the conditional attributes. To do so, we will first fine-tune GPT-2 (Radford et al., 2019) on the NewsAggregator dataset[2] which includes about 25,000 health, business, technology or entertainment news from March 10th, 2014 to August 10th, 2014. We will use a small version of GPT-2 composed of 12 blocks of decoder transformers pre-trained by huggingface (Wolf et al., 2020). Then, we will build an experiment protocol to ensure that our model generates text conditional on the attributes given as inputs.

2 Experiment Protocol

Below is the experiment protocol we followed in this project:

1. As mentioned above, we fine-tuned GPT-2 on the NewsAggregator dataset to be able to generate news conditionally on some attributes such as the category of the news, keywords and titles.

2. We generated about two hundred fake news conditioned on various attributes using either beam search or top-k method.

3. We then fitted a Latent Semantic Analysis on the generated news in order to get *news* embeddings based on the TF-IDF document matrix. We fitted a T-SNE on these *news* embeddings in order to vizualise the *news space* and identify clusters.

4. We fine-tuned a BERT (Devlin et al., 2019) classifier (pre-trained by huggingface) on the NewsAggregator dataset in order to predict the category given the news as input. Concretely, we freezed the weights of BERT and trained a classifier at the last layer. The aim is to compare the score obtained on the test set we built from the NewsAggregator dataset and the score obtained on the news we generated.

3 Results

We generated about 200 fake news in order to analyze the performances of our fine-tuned GPT-2. As mentioned above, we built news embeddings using a Latent Semantic Analysis on the TF-IDF matrix. More specifically, we chose to keep the first 128 eigenvalues which explain about 79% of the variance, and we then fitted a T-SNE on these embeddings.

In figure 1, it is possible to identify clusters according to the categories, meaning that the language model generates well-distinct news for dif-

[1] https://gitlab.com/matthieu_futeral/fake_news_generation

[2] https://archive.ics.uci.edu/ml/datasets/News+Aggregator

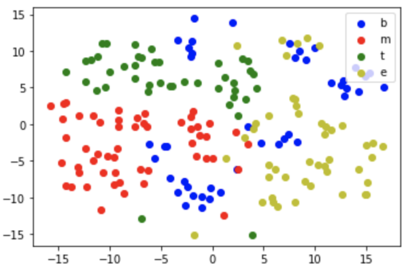

Figure 1: T-SNE on the fake news embeddings obtained with LSA

ferent categories. In addition, it is interesting to notice that the cluster ”health” (m) is close to the cluster ”technology” (t). Indeed, some news can deal with technology applied to health so it is not surprising that ”health” news are close to ”technology” news but it is interesting that the model succeeded in catching such a similarity. Furthermore, the cluster corresponding to the category ”entertainment” (e) is the further from the clusters ”health” (m) and ”technology” (t) which is understandable because these categories are supposed to be well distinct. Eventually, ”health” (m), ”technology” (t) or ”entertainment” news can have a business aspect, as well as ”business” (b) news can be related to health, technology or entertainment topics. Therefore, it is not surprising not to identify a compact business cluster because the separation is not supposed to be drastic between ”business” (b) news and others. Then, it seems that the model generates according to the requested category very well.

It is interesting to complete the same experiment on a few randomly selected news from the NewsAggregator dataset. In figure 2, it is possible to identify the same clusters but they are way more compact and there exist some outliers as well.

Then, we built an independent classifier to quantify how well it was possible to classify the category of the generated fake news in comparison with the test set of the NewsAggregator dataset. Concretely, as mentioned above, we fine-tuned a BERT classifier on the NewsAggregator dataset.

By comparing table 1 and 2, we can notice that the performance of the independent classifier on the generated news are slightly worse than the performance on the test set of the NewsAggregator dataset. It is therefore possible to conclude that the language model does not perfectly generate news according to the requested category... Indeed, if so, we would expect the performances of the independent classifier to be of the same order for both dataset but it is not the case (72% accuracy on the generated news against 77% on the test set), and the drop of performance can only be attributed to the language model.

	Precision	Recall	F1-score
bus.	0.87	0.45	0.60
tech.	0.52	0.86	0.65
ent.	0.85	0.85	0.85
health	0.83	0.74	0.78

Table 1: Classification report of the BERT classifier applied on the generated fake news

Eventually, we analyzed some generated news and determined whether it was plausible or not[3]. From text 1, it is possible to conclude that the model finds it hard to follow the title of the news. Indeed, it seems to generate an ”entertainment” news as requested but did not follow the title given as input: we wanted to generate a piece of news stating that Michael Jackson had been seen in Ibiza but there is no mention of that in the generated text despite the keyword ”alive” given as input. The same conclusion can be stated from text 2: even if the model is not supposed to know anything from the new coronavirus (as the training set is news from 2014), it could have generate a piece of news related to the fact that Bill Gates created a new coronavirus (a known disease as there were news about MERS in the training set), a fake news currently spread on the Internet. In text 3, the fake news is more related to the title given as input but it misses a part of it. The news never mentioned the fact that the cryptocurrency becomes the official currency in China as the title stated.

4 Discussion/Conclusion

Everything that has been reported needs to be qualified because our analysis is very dependent to the attributes we decided to use to generate fake news. It is however possible to draw some conclusions. GPT-2 is able to generate fake news conditional on a category but it fails to generate fake news according to more complex attributes such as titles and keywords that have more powerful meaning.

[3] Please refer to the appendix to read the fake news

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Thomas Wolf, Lysandre Debut, Victor Sanh, and Julien Chaumond. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Appendix

	Precision	Recall	F1-score
bus.	0.81	0.68	0.74
tech.	0.74	0.77	0.76
ent.	0.84	0.81	0.82
health	0.71	0.81	0.76

Table 2: Classification report of the BERT classifier applied on the test set of the NewsAggregator dataset

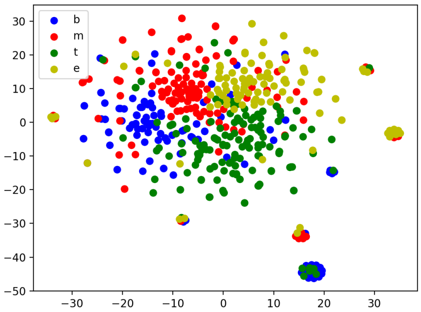

Figure 2: T-SNE on randomly chosen news from the NewsAggregator dataset, embeddings obtained with LSA

Michael J.Jackson is a famous man and singer who was born on May 5th of the same year as his father's birth certificate which he signed with an electronic device that read: "JIMMY JACKSON." The recording artist has been living out all over Miami since then but it appears to be getting closer for him now because according To Us Weekly, MJ's mother had recently moved into her new home at 3300 Nueces Ave., where she lives comfortably by herself while still wearing sunglasses [...]

Text 1: Generated news from the category "*entertainment*", the title "*Michael Jackson seen in Ibiza*", and the keywords "*popstar*" and "*alive*".

The first human case of the deadly Ebola virus has been reported after being carried out by an American scientist who had worked on it. Dr Kent Brantly was infected with West Africa's most lethal form when he tested positive for ebola last week and is now at home recovering from his infection but fears are that more cases will emerge soon as tests continue to confirm whether or not there were any links between him contracting noroviruses earlier this year [...]

Text 2: Generated news from the category "*health*", the title "*Bill Gates created new coronavirus in a secret lab, scientists said*", and the keyword "*disease*".

China's central bank has said it will soon issue a new cryptocurrency, known as "Bitcoin Cash. The Chinese government is expected to announce its decision in due course," Xinhua reported on Tuesday citing state media and other sources cited by Reuters news agency. "The announcement comes after two days when authorities announced that they would begin issuing virtual currencies for all citizen [...]

Text 3: Generated news from the category "*technology*", the title "*Bitcoin becomes the official currency of China*", and the keywords "*bitcoin*", "*china*" and "*currency*".