Published in Towards Data Science

rohola zandie Follow
Mar 10, 2021 · 11 min read · Listen

Topical Language Generation with Transformers

Controlling the large-language models generation capability is an important task that is needed for real-world usage.

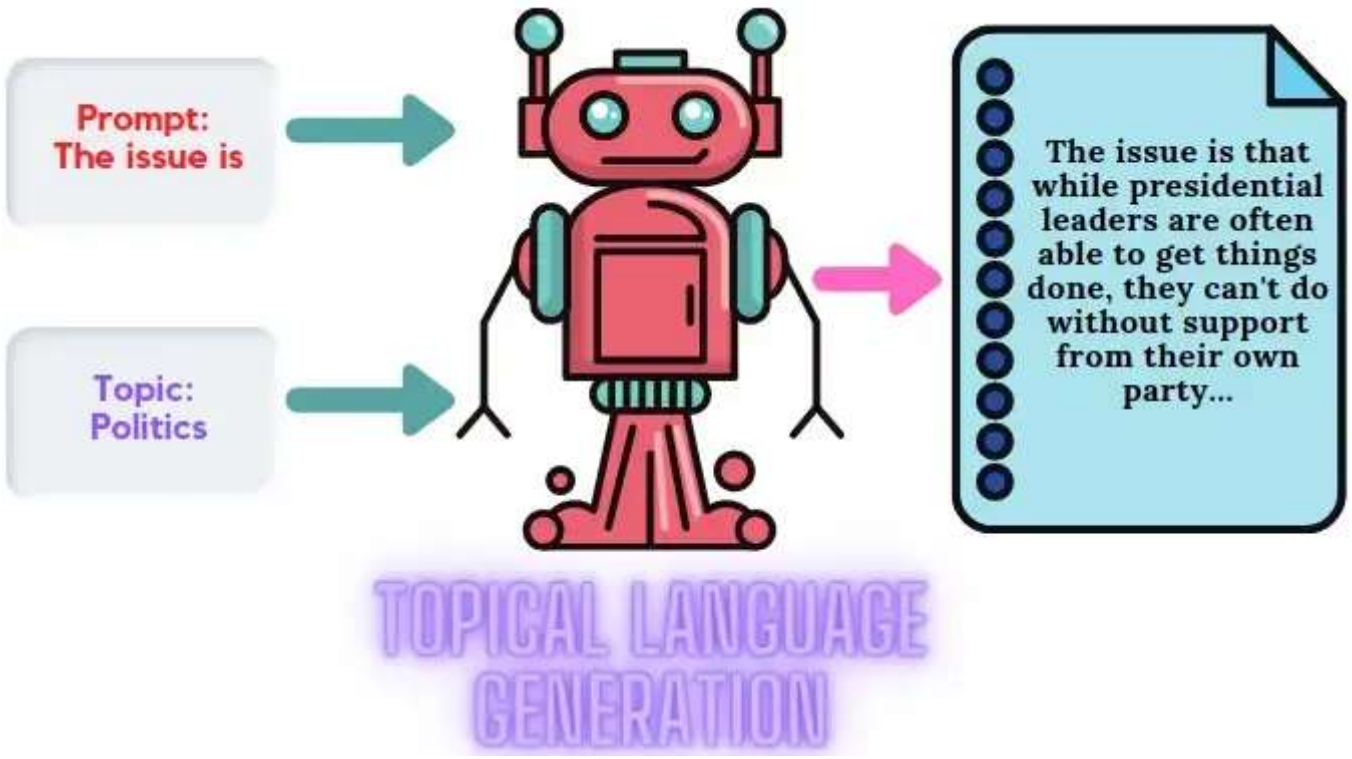


Image by Author

Full Paper

Codes

properties such as the topic, style, and sentiment is challenging and often requires significant changes to the model architecture or retraining and fine-tuning the model on new supervised data.

We introduce Topical Language Generation (TLG) by combining a pre-trained LM with topic modeling information. We cast the problem using Bayesian probability formulation with topic probabilities as a prior, LM probabilities as the likelihood, and topical language generation probability as the posterior. In learning the model, we derive the topic probability distribution from the user-provided document's natural structure. Furthermore, we extend our model by introducing new parameters and functions to influence the quantity of the topical features presented in the generated text. This feature would allow us to easily control the topical properties of the generated text.

Language modeling and decoding

The applications of language generation in NLP can be divided into two main categories: directed language generation and open-ended language generation. Directed language generation involves transforming input to output such as machine translation, summarization, etc. These approaches need some semantic alignment between the inputs and the outputs. On the other hand, open-ended language generation has much more freedom in the generation process because it does not need to be aligned with any output. The open-ended language generation has applications in conditional story generation, dialog systems, and predictive response generation. Even though there is more flexibility in choosing the next tokens compared to directed language generation, controlling the top-level features of the generated text is a desirable property that needs to be addressed and still is a challenging problem.

Given a sequence of m tokens x_1, .., x_m as the context, the problem of open-ended language generation can be formulated as finding the continuation x_{m+1}, ..., x_{m+n} with n tokens. In other words, if we consider the whole context plus continuation as following:


$$P(x_{1:m+n}) = \prod_{i=1}^{m+n} P(x_i|x_{<i})$$

The language modeling probability can be used with a *decoding strategy* to generate the next token for language generation. Finding the optimal continuation can be formulated as:

$$\hat{x}_{m+1:n} = \underset{x_{m+1:n}}{\operatorname{argmax}} P(x_{m+1:n}|x_{1:m})$$

Solving the above Equation is not tractable so practical decoding strategies use approximations to generate the next tokens. The most famous and widely used decoding strategies are greedy decoding and beam search methods. Greedy decoding selects the highest probability token at each time step, while the beam search keeps a set of hypotheses and then updates the tokens in the hypotheses as it goes through and decodes more tokens. These approaches are well suited for directed language generation, but they suffer from repetition, genericness, and degenerate continuations.

Both of these approaches are deterministic in the sense that they do not involve any random selection in their algorithms.

On the other hand, stochastic decoding methods sample from a model-dependent distribution q:

$$x_i \sim q(x_i|x_{<i}, p)$$

The simplest stochastic sampling consists of sampling from top-k probabilities, the use of constant k is problematic because in some contexts the probability distribution of


$$\sum_{x \in V^{(p)}} P(x|x_{<i}) \geq p$$

Then the resulting distribution which is based on the new vocabulary should be re-scaled to form a probability distribution. Under Nucleus Sampling, the number of plausible next tokens changes dynamically with the context and generated tokens. In this work, we use Nucleus Sampling as the base decoding technique and propose a new method to take into account topical knowledge about the tokens.

Topical Language Modeling

Given a list of K topics t = {1…K}, to control the outputs of the language model to follow a certain topic, at each generation step, we have to model the following probability distribution:

$$P(x_{1:m+n}|t_j) = \prod_{i=1}^{m+n} P(x_i|x_{<i}, t_j)$$

Compared to the previous Equation, the only difference is that it is conditioned on the topic t_j. To create the right-hand side of Equation 6, we change the last layer of the network that creates the logits.

Here, we adopt the GPT transformer architecture. If S is the last layer we use softmax to get the final probabilities:

$$P(x_i|x_{<i}) = \frac{\exp(S(x_i|x_{<i}))}{\sum_z \exp(S(z|x_{<i}))}$$

$$\sum_z P(z|x_{<i})P(t_j|z, x_{<i})$$

Because in topic modeling, documents are treated as bag of words we can also assume that the probability of the topic for each token is independent of the previously generated tokens. Based on this assumption we have:

$$P(t_j|x_i, x_{<i}) = P(t_j|x_i)$$

Now, assuming that we have P(t_j|x_i), then using Equation 10 we can prove that the conditions topical language model can be written as:

$$P(x_i|x_{<i}, t_j) = \frac{\exp(S(x_i|x_{<i}) + \log P(t_j|x_i))}{\sum_z \exp(S(z|x_{<i}) + \log P(t_j|z))}$$

For complete proof refer to the paper.

Topic modeling

Topic modeling algorithms automatically extract topics from a collection of textual data. They are based on statistical unsupervised models that discover the themes running through documents. We use two main algorithms in topic modeling.

- LDA (Latent Dirichlet Allocation): The basic idea behind LDA is that in a collection of documents, every document has multiple topics and each topic has a probability distribution. Moreover, each topic has a distribution over vocabulary. For example, a document can be on the topics of “Football”, “News” and “America” and the topic of “Football” can contain words including “NFL”, “Football”, “teams” with a higher probability compared to other words. Given a collection of M documents with vocabulary V, we can fix the number of topics to be K. In LDA, the probabilities of topics per documents and topic for tokens can be summarized in matrix forms, $\theta_{M\times K}$ and $\phi_{K\times|V|}$ respectively. After the learning, we have the distributions of

- LSI (Latent Semantic Indexing): LSI is the application of the singular value decomposition method to the word-document matrix, with rows and columns representing the words and documents, respectively. Let X_|V|×M be the token-document matrix such that X_{i,j} is the occurrence of token i in document j, then singular value decomposition can be used to find the low-rank approximation:

$$\hat{\mathbf{X}}_{|V|\times M} = \mathbf{U}_{|V|\times M}\mathbf{\Sigma}_{M\times M}\mathbf{V}^{T}_{M\times M}$$

After the decomposition, U still has the same number of rows as tokens but has fewer columns that represent latent space that is usually interpreted as “topics”. So, normalizing U gives us the scores of each token per topic. We can use this score for the probability of topic j for each token i in the vocabulary:

$$P(t_j|x_i) = \frac{\mathbf{U}^T[j,:]}{\|\mathbf{U}^T[j,:]\|}[i]$$

Controllable Generation Methods

The conditional topical language model in the equation above gives us a token generation that is conditioned on a specific topic but we cannot control the amount of the influence.

1- Adding topical parameter and logit threshold: adding the term log(P(t_j|x_i)) directly to the actual logit from the model can deteriorate the fluency of generated text in some cases. We propose two methods to alleviate this problem. We introduce a new parameter γ to control the influence of topical distribution:

$$P(x_i|x_{<i}, t_j) = \text{softmax}(S(x_i|x_{<i}) + \gamma \log(P(t_j|x_i)))$$

modeling.

The other approach is to cut the log probabilities of the topic with a threshold. The lower values of S correspond to tokens that the model gives very low probabilities and we do not want to change them because it introduces unwanted tokens and diminishes the fluency. In Equation above, we only keep log(P(t_j|x_i))for all the values of S that are larger than threshold.

$$\text{logprob}(i) = \begin{cases} \log\left(P\left(t_j \mid x_i\right)\right) & \mathrm{S}\left(x_i \mid x_{<i}\right) > threshold \\ 0 & \text{otherwise} \end{cases}$$

and log prob used in the following equation:

$$P(x_i|x_{<i}, t_j) = \text{softmax}(S(x_i|x_{<i}) + \gamma\ \text{logprob}(i))$$

ower values of threshold correlate with more on-topic text generation because we change more tokens from the original model by log(P(t_j|x_i)).

2 -Using α-entmax instead of softmax: The problem with the softmax function is that it gives non-zero probabilities to a lot of unnecessary and implausible tokens. The softmax function is dense because it is proportional to exp function and can never give exactly zero probabilities at the output. We useα-entmax instead to create more sparse probabilities that are less prone to degenerate text. α-entmax is defined as

$$\alpha\text{-entmax}(\mathbf{z}) := \underset{p \in \Delta^{|V|-1}}{\text{argmax}} \{\mathbf{p}^T \mathbf{z} + H_\alpha^T(\mathbf{p})\}$$

where Δ|V|−1:={p∈IR|V|−1,∑ipi=1}is the probability simplex, and for α≥1, HTα(p) is the

$$H_\alpha^T(\mathbf{p}) = \begin{cases} \alpha(\alpha-1)\sum_j (p_j - \\ -\sum_j p_j \log p_j & \alpha = 1 \end{cases}$$

α-entmax is the generalized form of the softmax function. In particular, for α=1 it exactly reduces to the softmax function and as α increases, the sparsity in the output probabilities continuously increases. Here we are specifically interested in α=2 which results in sparsemax:

$$\text{sparsemax}(\mathbf{z}) = \underset{\mathbf{p}\in\Delta^{|V|-1}}{\text{argmin}} \|\mathbf{p}-\mathbf{z}\|^2$$

Unlike the softmax function, sparsemax can assign zero probabilities.

3-Adding temperature and repetition penalty parameters: We need to make some changes to the base nucleus sampling to control the base distribution flatness and prevent it from generating repetitive words. We denote the final logit after the above changes as ui. Given a temperature, repetition penalty r and the list of generated tokens g, the final probability distribution for sampling is:

$$P(x_i|x_{<i}, t_j) = \text{softmax}(u_i/(T.R_g(x_i)))$$

when T→0, the sampling reduces to greedy sampling; while if T→∞ the distribution becomes flatter and more random. The penalized sampling discourages drawing already generated tokens.

Topical Text Generation with Different Topics

One of the biggest benefits of TLG is that it can be used with different language models without any retraining or fine-tuning of the base model, however, to generate topical texts we need to have topics extracted from a text corpus. For training the topic

raw text corpus.

In this experiment, a fixed neutral prompt has been used to make sure the model is not conditioned on the few initial tokens. The results in the table below show that after selecting a topic from the topic modeling output, the model can create long, coherent, and fluent text continuation without manually injecting extra knowledge from other resources or through training on labeled datasets.

Football	a shame because we've got so many young people who have been playing for us and they're all going out there fighting, but I don't know if you can imagine what the impact will be on them." He said each player would need professional training before being able go back into regular contact with any of his teammates behind teams having no official team or club affiliation between their clubs Football League!
Politics	The issue is that while presidential leaders are often able to get things done, they can't do them without the support of their own party members." In fact state legislatures have been reluctant to enact any kind for decades due largely government-initiated "party politics," which has led some states such as New York into a political crisis over how presidential candidates should approach issues like abortion and gay rights according in part those concerns about what it means given current trends between parties may be more important!
Media	The issue is that there will not be television coverage of news events in Russia. "We are going to have a lot less media," said three people familiar with what set up telecommunication services for the event. last month—TV medium companies like Vyldo, and Tmall Television & Radio transmitting their own content through its network, which has been shut down since May after being accused earlier of violating Russian law about broadcasting political messages on radio transmission platforms.
Physics	The issue is that when information about such radio waves are transmitted, the frequency fields of electromagnetic radiation can be measured."We have a lot technology to do," Aquo said. He added energy-threshold measurements could help scientists understand current conditions and how these signals interact with electrical circuits through sound modulating devices using amplitude oscillations and phase alternating conductor (A&P), "which allows pulse width modulation." When this happens back during space transmission medium properties formating an interference pattern.
Transportation	The issue is that car parts are not cheap because they can be bought at a much lower price than the original. I have seen many people who buy cars from Ford and say when you get them, almost all were sold for $100 US or less after having to pay more in taxes rather than buying new ones with higher prices (and I'm sure there would always depend on where vehicle was purchased). This has been true of most other vehicles since it started being used as an alternative fuel source!

Image by Author

Effects of Hyperparamters on TLG

In our proposed approach, we can useγandthresholdsas knob parameters to control the amount of topic influence on the language generation process. More specifically, based on Equation 27higher values of gamma will result in more on-topic results. Also, lower values of the threshold are correlated with more on-topic language generation

due to the fact that thresholding can easily cut off the probabilities that are related to function tokens (like stop words) in the vocabulary which hurts the fluency of the model. Fig below demonstrates the language generation on a fixed topic (football) with different values ofγandthreshold. To show how much each token accounts for the topic we use color-coding in which stronger colors show more on-topic words. We skipped the last stage of decoding. This is why the individual tokens from Byte Pair Encoding (BPE) tokenization can be seen.

president In im el Sports team sports leader Gael ic rugby professional teams regular season and international s league , NFL FC Conference each year have been asked over six years - in the last two decades between 2000 & 2010 offensive tackle played at least one game of Rugby League (NFL). The current rules allow only those who play !

$$\gamma = 20, threshold = -95$$

The issue is that some people are not aware how many different kinds of games you can play on each platform - and they 're all very similar ," he said . Ċ , played by Australian football ers in the 1980 s after playing for Australia against New Zealand Football teams between 1982 - 85 . In addition to being a great player himself he was also known as rugby league star when called up from North America at age 17 during World Cup qualifying matches with England team which won 2 out 3 League One titles (!

$$\gamma = 1, threshold = -95$$

The issue is that some football players are not allowed to wear head d resses . Ċ " I think it 's a very important thing when you 're playing for your country ," Football National Committee president In im el Un ic said Sunday after rugby league was banned during NFL season in the United States over professional athletes wearing teams ' uniforms , including those of their national team sports champions each week between January and March 17). Ċ NFL officials have been trying six months since then first banning all female members of regular - !

$$\gamma = 10, threshold = -95$$

The issue is usually between football teams playing professional rugby league games when each team sports regular season Football League National Conference champions Ċ A FC North America Super Bowl winners NFL AFC division wild card NFC playoffs conference tournament Sunday February 23 16 : 45 17 : 00 32 degrees Canadian soccer leagues played four major regional world title runs September 18 6 pm Australian Sports Un im el ic word association held December 19 1 week late ball play rules six different offensive codes divided equally early goal kicking score first level single line advance highest form !

$$\gamma = 5, threshold = -150$$

The issue is that the government has not been able to get a clear picture of how much money it will spend on its own infrastructure . Ċ , which includes roads and bridges - would be more expensive than other parts of National Capital Region (N RC). The cost for these projects could rise as well if they are built in areas where there was no traffic congestion or were under construction at regular intervals during peak hours such days when people travel by car rather then using public transport services like bus service from their !

$$\gamma = 5, threshold = -50$$

The language generation is the task of generating the next token conditioned on the previously generated tokens. The probability distribution of the next token in the base language models is flatter in some token positions and more peaked at some other token positions. For example, given the prompt of “The issue is that” there are plenty of possible next tokens compared to the next token of a prompt like “It is focused” which is almost always “on”. This property of language models gives us the flexibility to meddle in the generation process and steer it towards desired tokens when the probability distribution is flatter.

The concept of flat or peaked distribution can be easily measured in terms of the entropy of the distribution. In Figures a and b we compare the entropy of the base model (token entropy) with the posterior probability distribution from Equation 20 as the total entropy. Higher entropy for the base model in one position is a sign of its capability to sample from a large set of potential tokens with almost equal probabilities but in our conditional language modeling, we want to restrict that set to a smaller set that conforms with the chosen topic. Therefore, in almost all cases, the entropy of the TLG model drops significantly compared to the base model. We can observe the differences are larger for the tokens that represent the topic (like teams, football, culture and, music) and smaller for function tokens (like stop words that do not play any role in different topics).

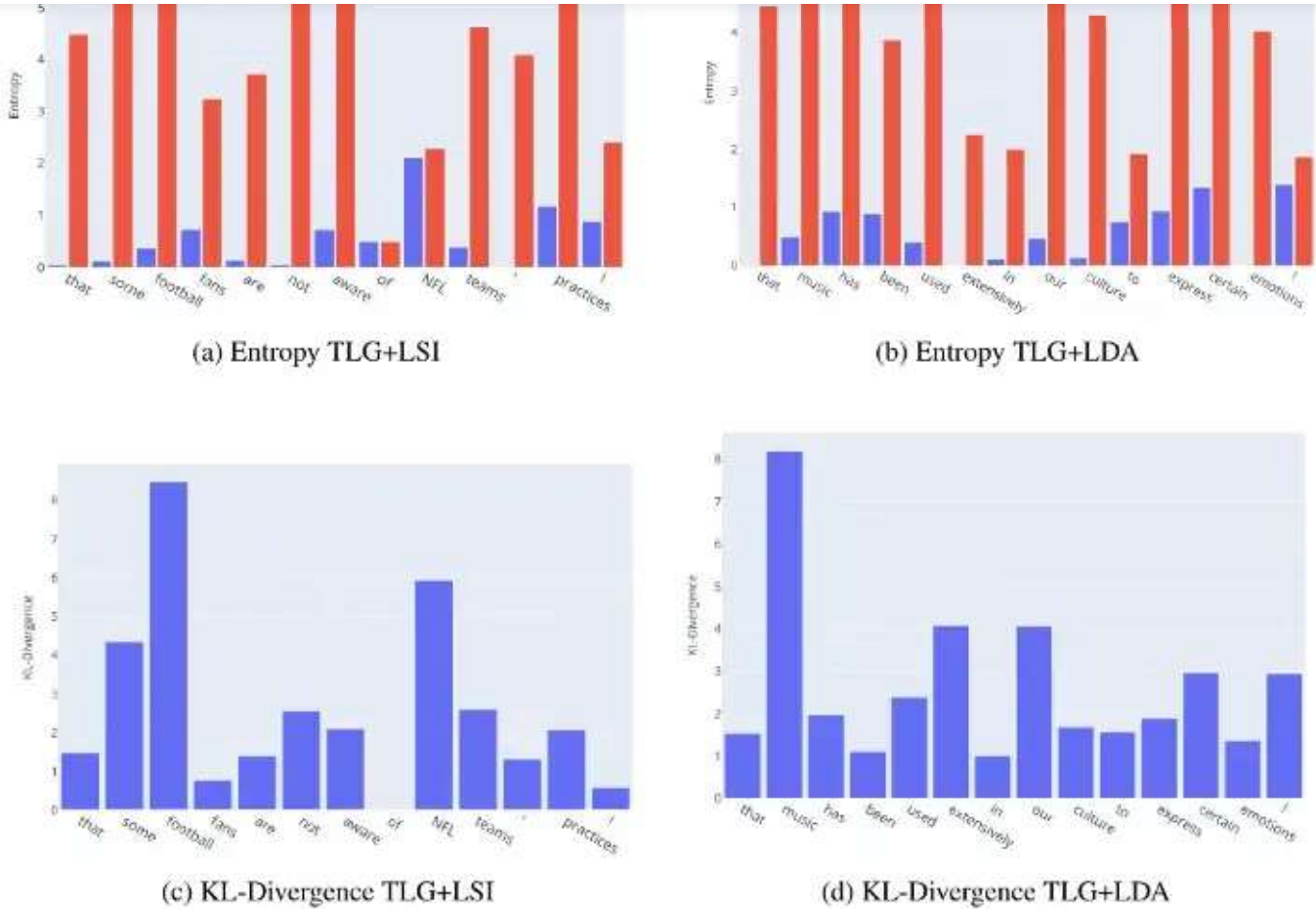
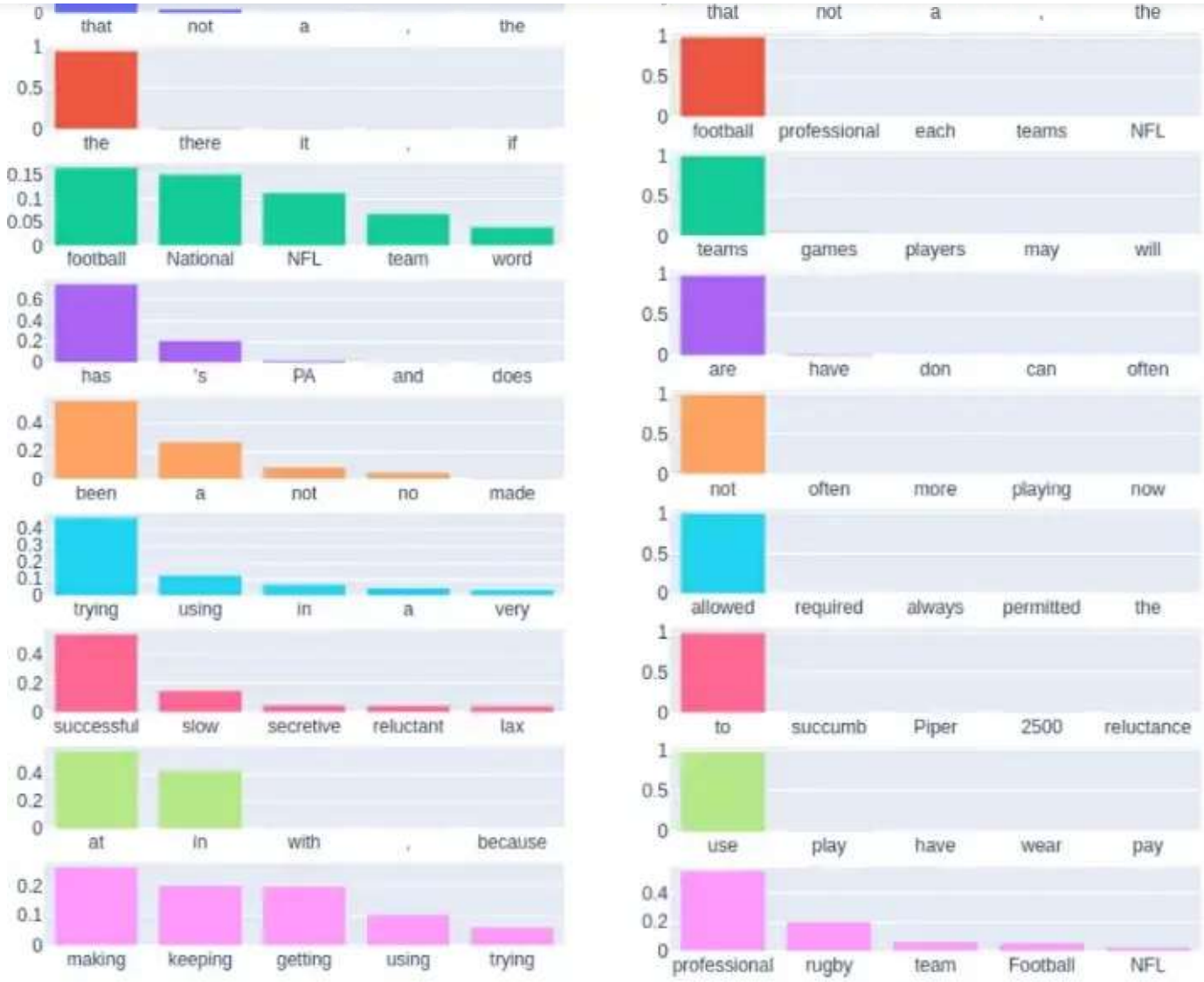


Image by Author

Another interesting observation is how the prior distribution that was extracted from topic modeling forces the language model to choose the topical tokens. The top-5 most likely tokens in a generation process are depicted in Figure 4. For the topic of football, the top-5 candidate tokens chosen by the model are compatible with the chosen topic.



(a) Softmax TLG+LSI (b) Sparsemax TLG+LDA

Image by Author

Graphical User Interface

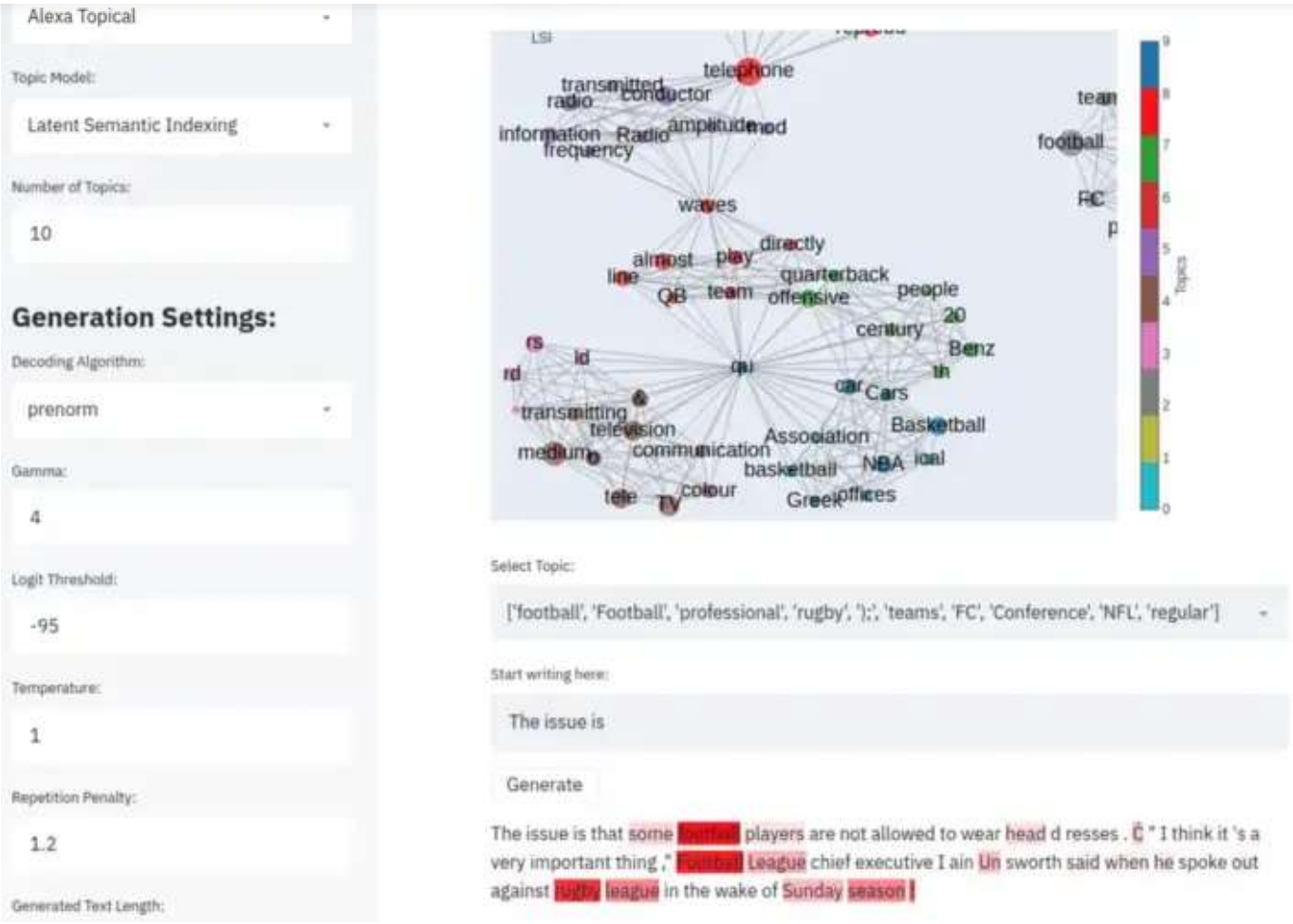


Image by Author

We also provide the GUI as a playground for users to work with the TLG. On the left panel, you can control the dataset, topic model, number of topics, and other generation settings. The playground gives you a graph plot which is a novel representation of the topics and how they are related to each other. Then you can choose the topic of interest and choose a prompt and finally hit the generate button to get the topical text.