Preference-based scoring model for political speech emotionality

Evžen Wybitul
Data Science, ETH Zurich
wybitul.evzen@gmail.com

I. Introduction

Large language models (LLMs) like Gemma [1] or Llama [2] excel at capturing nuanced contextual information, potentially leading to more accurate and insightful analyses in social science research. However, creating the necessary training data often requires time-consuming and expensive expert annotation, making it impractical for many research projects. This paper explores a novel approach to overcome this data scarcity problem, with implications for a wide range of social science applications.

We propose a two-step methodology: first, utilizing a powerful model to generate a large dataset of pairwise example comparisons, and then using this dataset to train a scoring model that quantifies the information relevant to the particular study. For the second step, we employ a preference learning approach to achieve more robust results, a technique recently popularized by reinforcement learning from human feedback (RLHF) [3].

To demonstrate our method's potential, we apply it to analyzing emotional content in political speeches, building upon the work of G. Gennaro and E. Ash [4] in *Emotion and Reason in Political Language*. Their study used word-level embeddings to measure speech emotionality, an approach that, while effective, has inherent limitations in capturing context-dependent emotional content. For instance, the sentence "They're taking our children's future" might appear neutral when analyzed solely at the word level, despite its clear emotional appeal when considered holistically. Thus, emotionality analysis serves as an ideal example of where more contextual methods could yield significant benefits.

We showcase our method's applicability by replicating and extending one of the main findings from G. Gennaro and E. Ash [4]: the evolution of speech emotionality over time. Our results (see Figure 1) not only confirm their observations, such as the increasing emotionality in political speeches over time and its correlation with major world events, but also uncover additional insights, including a peak in emotional content around the Korean and Vietnam wars era.

While this single application doesn't conclusively prove our approach's universal applicability, it offers compelling evidence of its potential. If successful, this methodology could significantly lower the barriers to applying advanced NLP techniques in social sciences, enabling researchers to gain insights into a wide array of problems previously challenging to address due to data limitations.

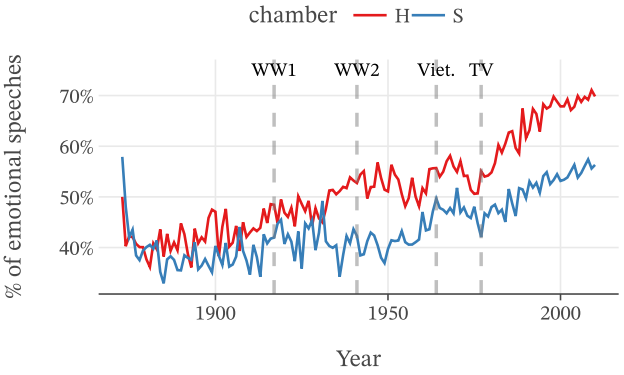

Figure 1: The percentage of emotional speeches in each year based on 150,000 pairwise speech comparisons. Lines mark important events: 1917 WW1, 1941 WW2, 1964 Vietnam war, 1977 C-SPAN founded. **Blue: Senate**, **Red: House**

In the following sections, we detail our methodology, present our findings, and discuss the broader implications of this approach for future research in social sciences. Our work aims to provide a stepping stone towards more widespread adoption of contextual models in social science research, potentially unlocking new avenues of inquiry and analysis.

II. Related Work

This study builds upon the work of G. Gennaro and E. Ash [4], who introduced a method to quantify emotionality in political speeches (hereafter referred to as the "original study"). Their approach involves classifying each speech as more emotional or more cognitive by aggregating the emotionality scores of its constituent words. These scores are computed by comparing word embeddings to vectors of words from the "emotion" and "cognition" categories, which are adapted from the Linguistic Inquiry and Word Count (LIWC) dictionary.

Our work shares the same goal as the original study: quantifying the emotionality of political speeches. However, we employ a different method based on contextual transformer

models trained on a set of labeled pairwise comparisons of speeches. This approach is similar in spirit to the work of P. Y. Wu, J. Nagler, J. A. Tucker, and S. Messing [5], who used GPT-3.5 to perform pairwise comparisons between congressional politicians based on various ideological metrics, such as party alignment and opinion on gun control laws.

Our method differs from that of P. Y. Wu, J. Nagler, J. A. Tucker, and S. Messing [5] in two key ways. First, due to the larger size of our dataset, we cannot feasibly label all pairs using GPT. Instead, we label a small subset and use it to train a more computationally efficient scoring model, which we then apply to label all the speeches. Second, instead of relying on the model's factual knowledge (e.g., asking "which of the following two politicians is the more emotional one?"), we directly provide the model with two speeches and let it decide based on the content at hand.

III. Dataset

Due to the original dataset used by G. Gennaro and E. Ash [4] being locked behind a paywall, we utilize a subset of the dataset presented by M. Gentzkow, J. M. Shapiro, and M. Taddy [6]. This subset contains transcripts from U.S. House and Senate speeches spanning from 1873 to 2011, as opposed to the original study's dataset which covered 1858 to 2014. To facilitate future research, we converted the dataset to the Parquet format and re-published both the original[1] and preprocessed[2] versions on HuggingFace.

It is crucial to note that the dataset was obtained through automated OCR on scanned documents, resulting in transcripts containing numerous typos and other errors. This issue is likely mitigated in the full licensed version of the dataset in the original study. An example of a representative excerpt from a speech in our dataset is provided in Listing 1.

In line with the original study, we filter out purely procedural speeches containing amendments. Furthermore, we only consider speeches with roughly between 64 and 1024 tokens, resulting in a total of 5,107,606 speeches. The number of speeches per year in this processed dataset is illustrated in Figure 2.

```
There is sonr co1nflict et ae e trnir of tire
corion bods ril tir registered bonds. which rnkes
it necessary fr the Urite Stares to express its
ehie whether it will ray- err tire lot of Jarniry
or not.
```

,

Listing 1: A representative excerpt from a speech taken verbatim from our dataset.

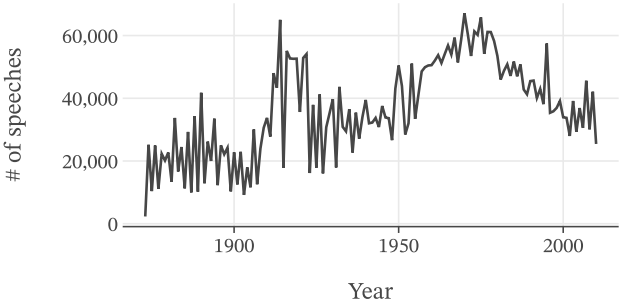

Figure 2: The number of speeches per year in the subset of the dataset we work with.

IV. Methods

This section outlines the three-step methodology employed in this study. For implementation details, please refer to the repository.[3] Our goal is to create a model which takes a single speech on input and produces a score that corresponds to its emotionality. For practical reasons detailed below, instead of performing direct supervised learning, we instead obtain a pairwise speech comparison dataset, and train the scoring model in a preference-based manner, similarly to how the reward model is trained in reinforcement learning with human feedback (RLHF) [3].

A. Step 1: Obtaining a Pairwise Emotionality Dataset

Due to the lack of labeled emotionality datasets for political speeches, we created our own. Unlike P. Y. Wu, J. Nagler, J. A. Tucker, and S. Messing [5], who work with a few thousand of pairs, our dataset contains >25 trillion unique pairs, making it infeasible to label all of them. Instead, we randomly selected 150,000 pairs for labeling.[4]

We used GPT-4o mini (henceforth GPT) for labeling instead of manual annotation. This approach is practical, saving human labor, and is partially based on current results showing that GPT-family models have some ability to understand emotions [7]. In a previous iteration of experiments before the release of the cheap *mini* version, we were limited to labeling only 5,000 pairs instead of 150,000, which made the results much worse.

Instead of directly asking GPT to assign a score to a single example, we provide it with pairs of speeches and ask it to determine which one is more emotional. This pairwise comparison technique has been shown to yield more stable and consistent results and is now standard practice [8], [3].

[1]https://huggingface.co/datasets/Eugleo/us-congressional-speeches

[2]https://huggingface.co/datasets/Eugleo/us-congressional-speeches-subset

[3]https://github.com/Eugleo/emotion-in-political-language

[4]This cost around $15 using the OpenAI Batch API. Labeling all pairs would come to over $2B, which is more than the yearly budget of Czech Republic.

To be more specific, we first ask GPT to think aloud about the emotionality of the two speeches, and then to provide a label 0, 1, or -1, depending on which speech it found to be more emotional — first, second, or neither, respectively. We do not use labeled as -1 during the subsequent training. The prompt we used can be seen in Listing 2.

```
Consider the following two excerpts from two
political speeches. Note that they contain spelling
mistakes and typos since they are were obtained by
OCR.

Excerpt 0: {excerpt_0}

Excerpt 1: {excerpt_1}

The goal is to judge which of these two, if any, is
more emotional. Don't take the topic itself into
account. Sometimes you will need to read between
the lines and infer the emotionality level, e.g.
in cases when the text is passive aggresive. Is
there any difference between Excerpt 1 and Excerpt
2 in terms of emotionality? If there is not, say
so.

Only after you're done with the analysis, end
the message with "result: 0" (without quotes) if
Excerpt 0 is the more emotional, or "result: 1"
if Excerpt 1 is more emotional. Use "result: -1"
if the result is unclear or there is no large
difference.
```

Listing 2: A representative excerpt from a speech taken verbatim from our dataset.

B. Step 2: Training the Emotionality Classifier

Our goal is to have a model that takes a single speech as input and produces an emotionality score. However, since our GPT-labeled dataset contains only binary comparisons of speech pairs, we cannot perform standard supervised learning. Instead, we use the Bradley-Terry model, introduced by R. A. Bradley and M. E. Terry [9] and widely used in applications like reinforcement learning from human feedback (RLHF) [3], to estimate scoring functions from pairwise comparisons.

1) *Bradley-Terry model:* Given two speeches x_1, x_2, we estimate their true emotionality scores f as $\hat{f}(x_1), \hat{f}(x_2)$, whose difference we interpret as the log odds of x_1 being more emotional than x_2:

$$\hat{f}(x_2) - \hat{f}(x_1) = \log \frac{P[f(x_2) > f(x_1)]}{1 - P[f(x_2) > f(x_1)]}. \qquad (1)$$

We fit $\hat{f}$ to minimize the cross-entropy between the true probabilities and the estimated ones, using our labeled dataset of ground-truth emotionality comparisons. For more details, refer to the original paper [9] or the RLHF paper [3].

2) *Implementation and hardware:* We use gemma-2-7b [1] as $\hat{f}$, due to its strong performance and compatibility with our hardware. We apply LoRA [10] with $r = 8, \alpha = 16$, and train the model for 1 epoch on 90% of the pairs using RewardTrainer from the HuggingFace trl[5] library, evaluating accuracy on the remaining 10%. Validation accuracy is used to verify the training process setup, but high validation accuracy does not guarantee generalization to unlabeled pairs. Training took around 14 hours on one Nvidia A40 GPU.

C. Labeling the speeches

Performing inference on the whole dataset would be too time-prohibitive, so we instead sample 50,000 speeches, or ~1% of the whole dataset, and use the fine-tuned gemma to compute the emotionality scores. We then take the average of these scores for each year and also report the standard error.

V. Results

TODO: Summary

A. Qualitative Assesment of GPT Pairwise Comparisons

We randomly sampled 150,000 pairs of speeches and obtained a score of 0, 1, or -1 denoting which speech in the pair, if any, is the more emotional one. As with our other datasets, we published the labeled pairwise dataset on HuggingFace.[6] You can see the distribution of labels in Table 1; the number of 0s and 1s is similar, as we would expect.

−1 (neither)	**0 (1st)**	**1 (2nd)**
1% (1420)	51% (76616)	48% (71964)

Table 1: Number of labels assigned by GPT to the pairs of speeches. Labels correspond to which speech of the pair GPT thought was more emotional.

Since we have no ground-truth dataset of emotional speeches to compare, we have to rely on a qualitative assessment of the results. We randomly inspected a few GPT judgements and they all appeared sensible. For an excerpt from a representative example, see Listing 3.

B. Evolution of Emotionality (Pairwise Comparisons)

Moreover, even without a properly trained scoring model, we can look at the ratio of how many speeches from a given year were classified as the more emotional ones from a pair to get a rough sense for the evolution of speech emotionality throughout the years. You can see the result in Figure 1.

We see that even without a scoring model, the results resemble the main figure from the original study. Of course,

[5]https://github.com/huggingface/trl

[6]https://huggingface.co/datasets/Eugleo/us-congressional-speeches-emotionality-pairs

Excerpt 0: `[...] The angel of death should have been called the devil of death on rail lines at Birkenau. the adjunct satellite camp to Auschwitz which the gentleman mentioned. [...]`

Excerpt 1: `[...] I know that the Armed Services Committee is always under a lightning rod. while other committees overspend and go over the budget figure with impunity. and nobody seems too uptight about that. [...]`

Decision: `Excerpt 0 exhibits a high level of emotionality through its passionate language, vivid imagery, and the invocation of grave historical events, specifically relating to the Holocaust. The speaker expresses strong feelings of anger and indignation towards actions and policies they believe lead to the loss of innocent life. The use of terms like "angel of death," "devil of death," and references to Dr. Mengele create an emotionally charged atmosphere. [...] result: 0`

Listing 3: An example of a speech pair from the speech pairwise dataset, together with the GPT written comparison and final answer.

we can not draw any causal findings from this plot, but it being similar to the one in the original study is some evidence for our method being roughly comparable to the original one.

Replicated results: We see that speech emotionality seems to be on the rise overall, with a few notable spikes (or in our case, bumps) around important events. As in the original study, the rise speeds up around the same time when C-SPAN was founded. We also see that the House is in general more emotional than the senate.

New results: Compared to the original study, we also observe a bump around the Korean and Vietnam wars, which is to be expected and confirms our method at least partially makes sense.

Other differences: In the original study, the curves of the Senate and the House seem to be largely aligned, and although the House is consistently more emotional, the overall development is similar for both. In our plot, however, emotionality in the Senate does not seem to rise much at all until the early 1960s. It is hard to interpret this; it could just be an artifact caused by our limited sampling. Also, since our dataset is only a subset of the original dataset, we are unable to observe the time span around the civil war.

C. Scoring Model Training

D. Evolution of Emotionality (Scoring Model)

VI. Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri tamen permagna accessio potest, si aliquod aeternum et infinitum impendere malum nobis opinemur. Quod idem licet transferre in voluptatem, ut postea variari voluptas distinguique possit, augeri amplificarique non possit. At etiam Athenis, ut e patre audiebam facete et urbane Stoicos irridente, statua est in quo a nobis philosophia defensa et collaudata est, cum id, quod maxime placeat, facere possimus, omnis voluptas assumenda est, omnis dolor repellendus. Temporibus autem quibusdam et.

One of the important open questions is how much overlap does there need to be between the individual pairs in the labeled pairwise dataset for the Bradley-Terry model to be still applicable. More research should also be done into describing the tradeoffs between this regime and using a LLM to score examples directly. In our specific case, it would also be interesting to see how much better the classifier would work if there weren't any typos in the dataset. It would also be very interesting to see — especially after we have a good emotionality scoring model — where the model places importance when it does the scoring. Employing a method such as sparse auto encoders or other mechanistic interpretability methods would be interesting.

References

[1] G. Team *et al.*, "Gemma: Open Models Based on Gemini Research and Technology." Accessed: Jul. 21, 2024. [Online]. Available: https://arxiv.org/abs/2403.08295v4

[2] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models." Accessed: Jul. 21, 2024. [Online]. Available: http://arxiv.org/abs/2307.09288

[3] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences." Accessed: Apr. 30, 2024. [Online]. Available: http://arxiv.org/abs/1706.03741

[4] G. Gennaro and E. Ash, "Emotion and Reason in Political Language," *The Economic Journal*, vol. 132, no. 643, pp. 1037–1059, Apr. 2022, doi: 10.1093/ej/ueab104.

[5] P. Y. Wu, J. Nagler, J. A. Tucker, and S. Messing, "Large Language Models Can Be Used to Estimate the Latent Positions of Politicians." Accessed: Apr. 30, 2024. [Online]. Available: http://arxiv.org/abs/2303.12057

[6] M. Gentzkow, J. M. Shapiro, and M. Taddy, "Congressional Record for the 43rd-114th Congresses: Parsed Speeches and Phrase Counts." Stanford Libraries, Palo Alto, CA, 16, 2018.

[7] Z. Lian *et al.*, "GPT-4V with Emotion: A Zero-shot Benchmark for Generalized Emotion Recognition." Accessed: Apr. 30, 2024. [Online]. Available: http://arxiv.org/abs/2312.04293

[8] Y. Wang *et al.*, "RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback." Accessed: Feb. 16, 2024. [Online]. Available: http://arxiv.org/abs/2402.03681

[9] R. A. Bradley and M. E. Terry, “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons,” *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952, doi: 10.2307/2334029.

[10] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models.” Accessed: Jul. 21, 2024. [Online]. Available: https://arxiv.org/abs/2106.09685v2