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[1] 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 usefulness 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.

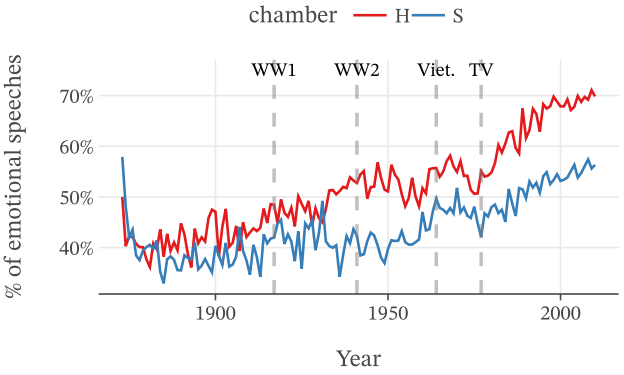

Figure 1: The percentage of emotional speeches in each year based on 150,000 pairwise speech comparisons. Dashed lines mark important events: 1917 WW1, 1941 WW2, 1964 Vietnam war, 1977 C-SPAN founded. **Blue: Senate**, **Red: House**

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.

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

[1]Our current best results are based directly on the pairwise comparison dataset (step 1), and do not utilize the scoring model. However, the training of the scoring model that we used to obtain these results has not fully concluded yet. Even in the case where the current scoring model will turn out not to work, even after full training, the first step at least seems to be shown useful by our paper.

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. Instead, we label a small subset and use it to train a more computationally efficient scoring model, which we can then apply for more comprehensive labelling, if needed. Second, instead of relying on the GPT's high-level factual knowledge (e.g., asking "which of the following two politicians is the more emotional one?"), we provide the model with two speeches and let it decide based on the actual raw content.

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[2] and preprocessed[3] 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.

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

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

```
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.

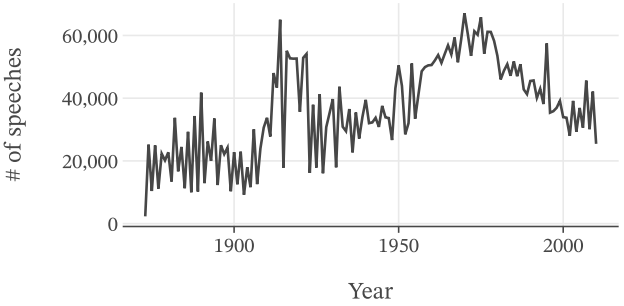

Figure 2: The number of speeches per year in the subset of the dataset we work with.

IV. Methods

This section outlines our two-step methodology. For implementation details, please refer to the repository.[4]

Our aim is to develop a model that takes a single speech as input and produces a score corresponding to its emotionality. Instead of direct supervised learning, we obtain a pairwise speech comparison dataset and train the scoring model using a preference-based approach, similar to the reward model training in RLHF [3].

A. Step 1: Obtaining a Pairwise Emotionality Dataset

Unlike P. Y. Wu, J. Nagler, J. A. Tucker, and S. Messing [5], who work with a few thousand pairs, our dataset contains >25 trillion unique pairs, making exhaustive labeling infeasible. We randomly selected 150,000 pairs for labeling. We employed GPT-4o mini (henceforth GPT) for labeling instead of manual annotation.[5] This approach is at least partially principled, as some results indicate GPT-family models have some ability to understand emotions [7].

Rather than asking GPT to assign a score to individual examples, we provide pairs of speeches and ask it to determine which is more emotional. This pairwise comparison technique yields more stable and consistent results and is now standard practice [8], [3]. We first ask GPT to analyze the emotionality of both speeches, then provide a label: 0, 1, or −1, depending on whether the first speech, second speech,

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

[5]This cost around $15 using the OpenAI Batch API. Labeling all pairs would exceed $2B, more than the yearly budget of the Czech Republic.

or neither is more emotional, respectively. Labels of -1 are not used in subsequent training. The prompt 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

We use the Bradley-Terry model, introduced by R. A. Bradley and M. E. Terry [9] and widely used in applications like RLHF [3], to estimate scoring functions from the pairwise comparisons obtained in step 1.

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

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

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

Implementation and hardware: We use gemma-2-9b [1] as $\hat{f}$, due to its strong performance and hardware compatibility. 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[6] library, evaluating accuracy on the remaining 10%. Validation accuracy verifies the training process setup but doesn't guarantee generalization to unlabeled pairs. Training took approximately 18 hours on one Nvidia A40 GPU.

C. Quantifying the Evolution of Emotionality

To gauge the evolution of speech emotionality over time, we examine the ratio of speeches from each year classified as more emotional within pairs. This proxy, based solely on step 1 and not involving any scoring model training, successfully replicates the result from the original study, as shown in Figure 1.

The scoring model allows us to compute an emotionality score for any speech, even those not included in the training dataset. As performing inference on all 5M speeches would be too costly, we sample 50,000 speeches (approximately 0.1% of the whole dataset) and use the fine-tuned gemma to compute their emotionality scores.

V. Results

We successfully replicated the findings of the original study and uncovered some new insights. Notably, these findings were discernible even from the raw dataset of pairwise comparisons, without employing a scoring model.

A. Qualitative Assesment of the Pairwise Labels

We sampled 150,000 speech pairs and obtained scores of 0, 1, or -1, indicating which speech in each pair, if any, was more emotional. The labeled pairwise dataset is available on HuggingFace.[7] The distribution of labels is shown in Table 1.

In the absence of a ground-truth dataset of emotional speeches, we rely on qualitative assessment of the results. Listing 3 provides an excerpt from a representative example.

B. Evolution of Emotionality (Pairwise Dataset)

To gauge the evolution of speech emotionality over time, we examined the ratio of speeches from each year classified as more emotional within pairs. The results are depicted in Figure 1.

Remarkably, even without a scoring model, the results closely resemble the main figure from the original study.

Replicated results: We observe an overall rise in speech emotionality, with notable spikes around important events. As in the original study, the rise accelerates around the time

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

Table 1: Distribution of labels assigned by GPT to speech pairs. Labels indicate which speech GPT deemed more emotional.

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

[7] 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: Example of a speech pair from the dataset, including GPT's comparison and final decision.

C-SPAN was founded. The House consistently demonstrates higher emotionality than the Senate.

New insights: Unlike the original study, we observe a bump around the Korean and Vietnam wars, which aligns with expectations.

Divergences: While the original study showed largely aligned curves for the Senate and House (with the House consistently more emotional), our plot suggests Senate emotionality remained relatively stable until the early 1960s. This discrepancy could be an artifact of our limited sampling. Additionally, due to our dataset being a subset of the original, we cannot observe the Civil War period.

C. Evolution of Emotionality (Scoring Model)

Note: Due to technical issues, the results in this section are based on a scoring model trained on only 20,000 examples instead of the full 150,000. We will update the report on the GitHub repository once the full training run is complete.

The model achieved an accuracy of 0.879 on the validation set. Compared to the proxy computed directly from the pairwise dataset, the scoring model does not recover all the structure that we are looking for, see Figure 3. It is hard to conclude whether this is due to its training not being finished yet, the pairwise dataset not being rich enough, or the model itself not being capable enough to generalize from the few pairwise examples to the whole dataset.

VI. Conclusion

In this study, we presented a novel approach to address the data scarcity problem when applying Large Language Models (LLMs) to social science research. Our methodology combines automatically labeled data with preference-based learning, offering a potential solution to the challenge of obtaining large, high-quality datasets in social sciences.

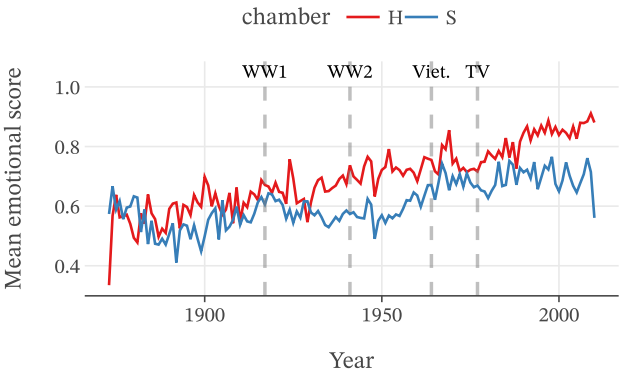

Figure 3: The mean score of emotional speeches in each year based on the trained scoring model. Dashed lines mark important events: 1917 WW1, 1941 WW2, 1964 Vietnam war, 1977 C-SPAN founded. **Blue: Senate**, **Red: House**

We demonstrated the viability of this approach by replicating a key finding from G. Gennaro and E. Ash [4] regarding the evolution of emotional content in U.S. congressional speeches. Our method not only successfully reproduced the original study's main results but also uncovered additional insights, such as identifying a peak in emotionality around the Korean and Vietnam wars era.

While these results are promising, we acknowledge that a single case study does not provide conclusive evidence for the universal applicability of our method. Further research is needed to assess its viability across different applications in social sciences. Future work could:

1. Attempt to replicate additional findings from the original study and explore the methodology's depth and breadth.
2. Investigate the optimal ratio of automatically obtained labels to human expert labels.
3. Incorporate ground truth data to better verify GPT's labeling accuracy and assess potential biases introduced by the models.
4. Explore the method's applicability to other social science domains and research questions.
5. Find the cause of the scoring model performing worse than the proxy based only on the pairwise dataset. *Note:* Work on this is in progress.

The abundance of open questions underscores that we are far from automating social science research. However, we believe this method could serve as a valuable tool for rapid hypothesis verification and idea iteration in social scientists'

workflows. It offers a cost-effective way to generate initial insights and guide more focused, expert-driven analyses.

In conclusion, while our approach shows promise in bridging the gap between advanced NLP techniques and social science research, it should be viewed as a complementary tool rather than a replacement for traditional methods. By enabling researchers to quickly explore large datasets and generate preliminary insights, this methodology could accelerate the pace of discovery and innovation in social sciences, opening new avenues for inquiry and analysis.

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