How and Why LLMs Rewrite Text: A Study in Media Bias Mitigation

•••

Neel Iyer

Guiding Question

How can we use generative technology to promote media fairness and objectivity?

- New technology like ChatGPT: part of a family of large language models (LLMs)
- Have unique potential in being able to comprehend and transform text
  - Utilize it for media transformation → utilization of technology for societal good

Media Bias Literature Review

Bias in news is a byproduct of having a media system in the first place and is perennial (Lowry 2008) and is typically generated through political framing to increase subjectivity (Lazaridou and Krestel 2016). This bias manifests itself through ‘subjectivity’ which is a quantifiable way to analyze bias’ presence (Aggarwal et al. 2020). Given the large majority of Americans that consume news, tackling this subjectivity is absolutely essential (Groeling 2013).

Large Language Models (LLMs) Literature Review

In context of bias mitigation, removing bias through eliminating words that contribute heavily to subjectivity has been explored with language models like BERT (Pryzant et al. 2020). But this failed to tackle the real issue → rewriting text to eliminate slant or political bias, but rather simply changing words on the surface.

Specific Research Question and Goals

- Can large language models that are generative in nature utilize a designed prompt to lower media subjectivity while preserving objective content?

Goals

- Transformation: subjective slant → objective facts
- Develop news datasets that serve as generalized representations for political media
- Understand how LLMs process text and interpret prompts → provide future avenues in prompt engineering

Inspiration

- Jan 6th led to political polarization
- Media political spectrums → source-source bias
- Subjectivity leads to a worsening of media quality and studies have shown can lead to citizens consuming less news as a whole (not helping citizen education)

Methodology: Data Collection

First, I selected 12 topics as media queries that would result in articles with contrasting views. I used Google News as my media aggregator to promote both relevance and public traction. Aggregating media over 50+ sources, I ended up scraping a total of 1098 articles under these 12 topics, with roughly equal distributions.

Methodology: Prompt Engineering

I used GPT3.5 as a testing LLM to prompt engineer a prompt for the purpose of rewriting to promote objectivity and tested it on sample cases. My final prompt was: “Rewrite text to be concise, fact driven, objective, neutral statements that aren't lengthy. They shouldn't hold any political affiliation, but be fact driven and short. You should have a formal, neutral tone, and not express any one view, but present both sides equally and fairly.”

Methodology: Subjectivity Benchmarking

Next, I had to create a benchmarking mechanism. Using a sentiment pipeline from TextBlob, a Python library, I was able to extract the subjectivity from each source. Here, subjectivity was calculated on a word based level, and multiplied by an intensity factor determined by word modifiers

Evaluating If Results are Statistically Significant

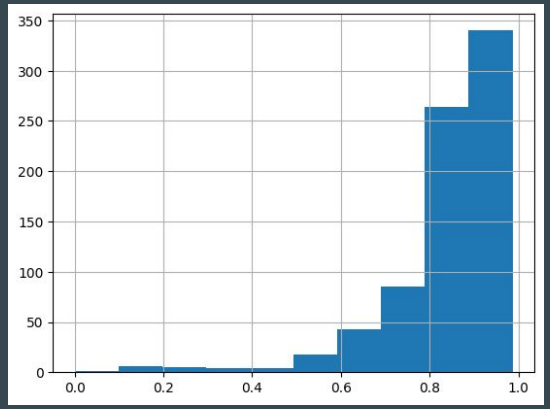
For our conclusions, we first ran a paired t-test to ensure our results were statistically significant (they were). We noticed on average a 10% decrease in subjectivity found from LLM rewritten text.

	Original	Rewritten
Mean	0.39725	0.35379
Variance	0.00102	0.00213
Observations	12	12
Hypothesized Mean Difference	0	
P(T<=t) one-tail 0.00028	0.00028	
t Critical one-tail	1.79588	

So we’ve shown LLMs can rewrite text to reduce bias. But how? What are the properties of rewritten text?

Is rewritten text similar? (Cosine Similarity)

I ran a benchmarking test to evaluate if the model really preserves the contents of the original text by comparing if the text vectors are close in distance. To do this, I utilized Spacy (a library's) largest English model to generate word-word embeddings for each document. Comparing the embeddings of the original document to the rewritten one with cosine similarity, we achieved a cosine similarity of 0.84, indicating results are highly similar.



Average Original-Rewritten Cosine Similarity
0.84

Evaluating Sentence Dissimilarities

We then compared the length of responses, finding that the model both decreased sentence length and average sentence subjectivity, indicating it removed most subjective and unnecessary content

	Original	Rewritten
Average Sentence Length (words)	26.822	20.868
Average Sentence Subjectivity	0.318	0.263

Clustering and Correlation

Lastly, we clustered our articles through K-means clustering, finding that for certain clusters, there was a moderately positive correlation between initial subjectivity and efficacy of model rewriting, displaying that our model works best on highly polarized/subjective datasets.

	orig_clusters	orig_sub	rew_sub	sub_diff
orig_clusters	1	0.1318	-0.100111	-0.354226
orig_sub	0.1318	1	0.670092	-0.44239
rew_sub	-0.100111	0.670092	1	0.366735
sub_diff	-0.354226	-0.44239	0.366735	1

orig_clusters	orig_sub
0	0.379567
1	0.399523
2	0.361345
3	0.616958
4	0.324675
5	0.424299

Final Recap and Implications

Thus, LLMs do have the potential to reduce subjectivity. This research can be used in 2 ways: to be able to quantify bias via rewriting the article and contrasting to evaluate whether bias was inherently present, and to provide users and consumers with the option to consume a more objective form of news that doesn't contain political slant.

Recommendations, Future Work, and a Possible Demo

Recommendations include having media outlets using the methodology outlined here and comparing benchmarked subjectivity until a percent difference of <2%. This will ensure that human written news is as objective as possible. Moreover, this can be extended in the future through having multiple prompts and analyzing how specific prompt engineering can impact the success rate of large language models.

Demo: website https://msec-powerofllms.streamlit.app/

Thanks!

...