Examining and Debiasing Political Bias in Web Search Results

Ronak Chaudhuri
rchau5@illinois.edu
University of Illinois
Urbana-Champaign
Illinois, USA

Ronje Roy
ronojoy2@illinois.edu
University of Illinois
Urbana-Champaign
Illinois, USA

Alex Neibart
neibart2@illinois.edu
University of Illinois
Urbana-Champaign
Illinois, USA

ABSTRACT

Political bias in web search results can significantly impact public perception and reinforce ideological echo chambers. In this paper, we investigate methods to quantify political bias in search engine outputs and propose a reranking algorithm that balances ideological diversity while maintaining relevance. Using existing datasets with ideological labels, we develop a bias scoring system based on entropy and KL divergence. Our results demonstrate that our approach effectively reduces bias without compromising the quality of search results. This work contributes to the development of fairer retrieval systems, with implications for news aggregators and general-purpose search engines.

1 INTRODUCTION

1.1 Problem Statement

Political bias in web search results can distort users' understanding of information. Search engines rank and display content in response to user queries. However, these algorithms may favor certain political perspectives, either due to inherent biases in the data or the algorithm itself. This can lead to biased information retrieval, influencing public opinion and reinforcing ideological silos.

1.2 Motivation

Addressing political bias in search engines is essential for maintaining balanced information access. Search engines play a major role in shaping how people perceive news and other content. When biased results dominate, they can skew public perception and foster polarized viewpoints. Mitigating bias helps ensure that users are exposed to a diverse range of perspectives, contributing to a more informed public discourse.

1.3 Research Question

Can we usefully quantify political bias in web search results? Furthermore, can we reduce this bias while preserving relevance and user satisfaction?

1.4 Significance

Mitigating political bias in search results has profound implications for promoting fairness and balance in information retrieval. Search engines influence millions of users daily, and biased content can reinforce echo chambers, amplifying ideological divides. Developing methods to identify and reduce bias can support a more neutral and informative web environment. This work contributes to the ongoing efforts to make search engines more accountable and transparent.

1.5 Novelty

Many previous studies have investigated political bias in search engines, examining how algorithms and data sources influence biased results. However, most of these studies do not directly quantify bias using labeled data. Our approach addresses this gap by using datasets that explicitly categorize articles by political affiliation (left, center, right). This allows us to measure bias in a clear and quantifiable way.

The core novelty of our work lies in our reranking algorithm, designed to reduce bias while maintaining relevance. Unlike existing methods that often compromise one for the other, our approach dynamically balances both. By incorporating topic-specific adjustments and adaptive parameter tuning, our reranking method ensures that diverse viewpoints are represented without sacrificing the quality of search results.

2 RELATED WORK

2.1 Literature Review

Research on political bias in search results has explored the roots and effects of biased information retrieval. Several studies have examined how search algorithms can inadvertently amplify specific political perspectives.

Bias in Political Search. Kulshrestha et al. [3] examined the distinction between data bias and algorithm bias in political searches on social media. They found that bias can emerge from both the content being indexed and the way algorithms rank search results. The study highlighted that differently phrased but similar queries can yield varying levels of bias, suggesting that both algorithmic design and data characteristics impact bias levels.

Search Engine Bias Assessment. Mowshowitz and Kawaguchi [4] discussed the inherent challenges in evaluating bias within search engines. They argued that bias is difficult to measure because determining a document's relevance is inherently subjective. They proposed measuring bias by collating search results from multiple engines to form an "ideal distribution" for comparison.

User Behavior and Bias. White [5] explored how user interactions can introduce or amplify bias in search results. The study found that users' own preferences and search patterns can create a feedback loop, where biased results influence subsequent queries. This dynamic interaction between user behavior and algorithmic bias complicates efforts to achieve neutral search outcomes.

Fairness in Rankings. Biega et al. [2] addressed bias from a fairness perspective, focusing on how search algorithms can inadvertently discriminate against certain groups. Their work introduced methods to balance attention across ranked results, aiming to maintain fairness without sacrificing relevance. This approach is valuable for mitigating discrimination while preserving user satisfaction.

2.1.1 Summary. As a whole, existing research has largely focused on quantifying bias, in terms of both political ideology and discrimination. In doing so, researchers have explored several different bias metrics, including difference from a theoretical ideal distribution of search results and means of breaking bias down into components related to different aspects of the search process. The literature also discusses the ways in which bias emerges, and how it is affected by factors such as user behavior, the underlying dataset, and the way in which queries are phrased.

2.2 Gap in Existing Research

While these studies provide valuable insights into the origins and effects of political bias in search results, they often fall short in directly quantifying bias using labeled data. Additionally, few existing approaches effectively balance the reduction of bias with the preservation of relevance. Our project fills this gap by using datasets with ideological labels to measure bias quantitatively and developing a reranking algorithm that maintains relevance while reducing bias.

3 METHODOLOGY

3.1 Dataset Collection

We used the AllSides [1] dataset to categorize news articles by political orientation—Left, Center, or Right. While this dataset provides a structured way to assess ideological balance, it is important to recognize that the labeling process itself can introduce bias. Understanding how AllSides assigns these labels helps us evaluate the reliability of our bias quantification methods.

AllSides [1] employs a multipronged approach to rate media bias:

- **Editorial Reviews**: A team with diverse political perspectives analyzes content to assess bias.
- **Blind Bias Surveys**: Participants from across the political spectrum rate content without knowing its source, helping to mitigate preconceived notions.
- **Third-Party Data and Independent Research**: AllSides incorporates external analyses and research to inform their ratings.

These methods aim to balance subjective judgments and reduce individual biases. However, no system is free from limitations. The reliance on human judgment means that some degree of subjectivity is inevitable. Additionally, the dynamic nature of media outlets can lead to shifts in bias over time, which may not be immediately reflected in the ratings.

In our project, we acknowledge these limitations and use the AllSides labels as a practical tool for measuring ideological skew. By being transparent about the dataset's origins and its potential biases, we aim to provide a nuanced analysis of political bias in search results.

3.1.1 Preprocessing Steps. After acquiring the dataset, we performed several preprocessing steps to ensure data quality and consistency:

- **Data Cleaning**: We removed duplicates and handled missing values to maintain the integrity of the dataset. Additionally, we filtered out non-political articles to focus on relevant content.
- **Text Normalization**: We converted all text to lowercase to eliminate case-related discrepancies. We also removed special characters, punctuation, and common stopwords to streamline the analysis.
- **Labeling**: We categorized articles based on the AllSides ideological tags (Left, Center, Right). Each article was tagged according to the bias rating provided in the dataset.

These preprocessing steps helped standardize the data, making it more suitable for subsequent analysis and model training. We documented the transformations applied at each stage to ensure reproducibility and transparency.

3.2 Classification

Our classification system aims to categorize political news articles based on their ideological orientation. We implemented two distinct models: a binary classification model and a three-way classification model. Both models use RoBERTa-based classifiers, but they differ in how they handle ideological categories and how they calculate neutrality scores. The goal is to assess which model performs better at reducing bias while maintaining relevance and ideological diversity.

3.2.1 Binary Model Classification. The binary model outputs a single score between 0 and 1, representing a spectrum between left-leaning and right-leaning. This score helps classify documents based on political leaning:

- If the score is less than 0.5, the document is classified as left-leaning.
- If the score is greater than 0.5, the document is classified as right-leaning.

We use a sigmoid activation function to perform binary classification. The results of the binary ROBERTA-based model are shown below (notably, accuracy was 0.72):

```
Predicted label distribution: Counter({np.int64(0): 1540, np.int64(1): 692})
True label distribution: Counter({0: 1287, 1: 945})

=== Full RoBERTa Binary Classification Report ===
              precision    recall  f1-score   support

        Left       0.72      0.86      0.78      1287
       Right       0.73      0.54      0.62       945

    accuracy                           0.72      2232
   macro avg       0.72      0.70      0.70      2232
weighted avg       0.72      0.72      0.71      2232


✅ Full RoBERTa binary classifier saved to ./roberta_binary_model_full
Text: The progressive movement advocates for climate cha...
Neutrality score: 0.773

Text: Traditional values and constitutional rights are f...
Neutrality score: 0.874

Text: A balanced approach considers both economic growth...
Neutrality score: 0.293
```

Figure 1: Binary classification results showing prediction distribution and neutrality scores from the RoBERTa model.

3.2.2 Three-way Model Classification. The three-way model outputs three probabilities representing the likelihood that a document is left-leaning, neutral, or right-leaning. This score similarly helps classify documents based on political leaning:

- If the score is less than 0.35, the document is classified as left-leaning.
- If the score is between 0.35 and 0.65, the document is classified as neutral.
- If the score is greater than 0.65, the document is classified as right-leaning.

This model uses a softmax activation function to calculate the probability distribution. Similarly to the binary model, we use the MLE class as our final classification. The results of the full ROBERTA-based model are shown below (notably, accuracy was 0.63):

```
=== RoBERTa 3-Way Classification Report ===
              precision    recall  f1-score   support

      Center       0.52      0.56      0.54       600
        Left       0.66      0.69      0.67      1287
       Right       0.63      0.55      0.59       945

    accuracy                           0.62      2832
   macro avg       0.60      0.60      0.60      2832
weighted avg       0.62      0.62      0.62      2832

Text: The progressive movement advocates for climate cha...
Neutrality score (Center class prob): 0.378

Text: Traditional values and constitutional rights are f...
Neutrality score (Center class prob): 0.090

Text: A balanced approach considers both economic growth...
Neutrality score (Center class prob): 0.113
```

Figure 2: Three-way classification results showing prediction distribution and neutrality scores from the RoBERTa model.

3.3 Baseline Information Retrieval System

The Baseline Information Retrieval (IR) system forms the foundation of our search functionality. Its primary purpose is to provide an initial ranking of search results based on relevance. We use this system as a benchmark to evaluate how well our bias-reducing methods perform. By establishing a reliable baseline, we can measure improvements brought by our reranking techniques.

3.3.1 IR Methods. To achieve this, we implemented two main retrieval methods: BM25 and an embedding-based retrieval system.

BM25 (Default Method):

- BM25 is widely used in information retrieval for its efficiency and effectiveness when ranking large collections of documents.
- It calculates relevance based on:
  - **Term frequency:** How often a term appears in a document.
  - **Inverse document frequency:** How unique a term is across the corpus.
  - **Document length normalization:** Adjusts for longer documents to prevent bias.
- We used the `rank_bm25` library, specifically the BM25Okapi implementation, to compute relevance scores.
- To make BM25 more suited for our task, we incorporated query-specific bias adjustments, allowing the model to account for variations in how bias might present across different topics.

Embedding-based Retrieval:

- This method leverages semantic representations rather than pure keyword matching.
- We used the `sentence-transformers` library and selected the `all-MiniLM-L6-v2` model to generate both document and query embeddings.
- Cosine similarity measures the closeness between query and document embeddings.
- This approach captures semantic relevance more effectively than traditional keyword-based methods.

3.3.2 Technical Setup. Our technical setup consists of two main steps: indexing and the search process. During indexing, we prepare documents by either tokenizing them (for BM25) or generating embeddings (for embedding-based retrieval). Tokenization involves converting text to lowercase, removing special characters, and eliminating common stopwords. This process ensures uniformity and reduces noise in the data, making it more consistent and manageable for retrieval.

The search process itself begins by checking whether embedding-based retrieval is enabled. If so, the system computes cosine similarities between the query embedding and precomputed document embeddings. If BM25 is chosen instead, the system tokenizes the query and retrieves scores from the BM25 index. To balance relevance and ideological diversity, the final ranking combines relevance scores from the selected method with bias scores.

3.3.3 Why BM25 is Suitable. BM25 is particularly well-suited for our project for several reasons. First, it is efficient, allowing it to process large datasets without significant computational overhead. Second, it is effective, consistently performing well in traditional information retrieval tasks. Third, it is flexible, allowing us to incorporate bias adjustments as needed. Finally, it is interpretable, with scores grounded in well-established statistical principles. This makes the results reliable and easier to analyze.

3.3.4 Enhanced Features. To further improve BM25, we integrated topic-specific bias weights. These weights address content-specific bias, for example, economic topics may exhibit different bias characteristics compared to social issues. We also developed a combined scoring method that merges relevance and bias scores. This method ensures that the final ranking maintains both relevance and ideological balance.

3.3.5 Significance. The Baseline IR system plays a critical role in our project. It sets the standard for relevance-based retrieval and provides a starting point to measure the effectiveness of our bias-reduction techniques. By combining traditional and modern retrieval methods, it helps us explore the trade-off between relevance and bias reduction comprehensively.

3.4 Bias Quantification

Our goal in bias quantification is to measure the extent of political bias present in search results. To achieve this, we use a range of metrics that evaluate both the diversity and balance of the retrieved documents. These metrics help us understand whether our methods effectively reduce bias while maintaining ideological diversity.

3.4.1 Scoring Methods. To assess bias, we primarily use two scoring methods: Entropy and KL Divergence.

Entropy measures ideological diversity within the search results. It quantifies how evenly distributed the ideological viewpoints are among the documents. Higher entropy values indicate a more balanced and diverse set of perspectives, while lower entropy values suggest a concentration around specific ideologies. In our context, maintaining higher entropy is desirable because it reflects a more varied representation of political views.

KL Divergence measures the deviation of the current distribution from a balanced, uniform distribution. A lower KL divergence value indicates that the distribution of viewpoints is closer to being balanced. Conversely, a high KL divergence suggests that one or more ideological categories dominate, leading to a skewed distribution. By calculating these metrics, we can quantify how well our models maintain neutrality and reduce ideological skew.

3.4.2 Neutrality Score Calculation. To calculate neutrality scores, we use RoBERTa-based classifiers in both our binary and three-way models.

The **Binary Model** outputs a single score between 0 and 1 using a sigmoid activation function. This score represents the probability of neutrality.

In contrast, the **Three-way Model** outputs three probabilities representing left-leaning, neutral, and right-leaning tendencies using a softmax activation function. The model uses the probability of the middle class (neutral) as the neutrality score. The same thresholds are applied to categorize the output.

This approach allows both models to systematically quantify the likelihood of a document being neutral, left-leaning, or right-leaning. By comparing the outputs from both models, we gain insights into how each model interprets neutrality and how they handle ideological diversity.

3.4.3 Lexicon-Based Bias Analysis. In addition to using machine learning-based models, we also incorporate a lexicon-based analysis. This approach uses predefined lists of politically charged terms to calculate bias scores.

For **left-leaning terms**, we include words like "progressive", "diversity", "equity", "climate", and "justice". These terms are commonly associated with liberal viewpoints. For **right-leaning terms**, we use words like "patriot", "freedom", "taxes", "border", and "constitution", which are frequently linked to conservative ideologies.

We calculate bias by measuring the frequency of these terms within documents. Each term is assigned a weight:

- Left: 1.0
- Right: 1.0
- Neutral: 0.5

This method helps detect whether the language of a document leans towards a particular political orientation.

3.4.4 Bias Metrics Calculation. To quantify bias, we compute several metrics, including basic statistics, viewpoint entropy, KL divergence, and position-weighted bias.

We start by calculating **basic statistics**, such as the average and standard deviation of neutrality scores. These metrics help assess the central tendency and variability of the ideological distribution.

Next, we calculate **viewpoint entropy** by binning documents into left-leaning, neutral, or right-leaning categories and normalizing their frequencies. Higher entropy values indicate a balanced set of viewpoints, while lower values suggest ideological concentration.

We also compute **KL divergence** by comparing the current distribution of ideological labels with a uniform distribution. A higher KL divergence indicates that the current distribution is significantly skewed, while a lower value suggests that the viewpoints are more evenly distributed.

To account for ranking effects, we calculate **position-weighted bias**. This metric applies position weights similar to DCG (Discounted Cumulative Gain) to prioritize higher-ranked results. We then calculate a weighted average of neutrality scores based on their positions, ensuring that top results influence the overall bias score more significantly.

3.4.5 Topic-Specific Bias Weights. Bias levels can vary significantly depending on the topic. To address this, we apply topic-specific bias weights. For example, in economic topics, left-leaning terms may receive a lower weight compared to right-leaning terms. In healthcare topics, the pattern might be reversed. This approach helps account for inherent biases that are specific to the content area, allowing for more accurate bias measurement.

3.4.6 Diversity Calculation. We also calculate diversity between documents to understand how varied the perspectives are. This involves comparing neutrality scores of document pairs to assess ideological differences. We use a sigmoid function to capture nuanced differences and add a complementary bonus when contrasting pairs (such as left-right combinations) are detected. This method ensures that documents representing opposite viewpoints are recognized and valued for their diversity.

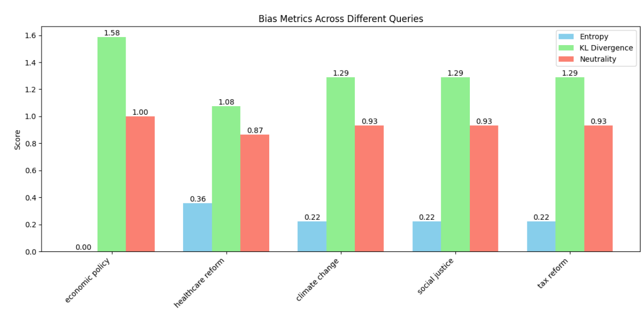

Figure 3: Bias Metrics Across Different Queries. The chart displays the Entropy, KL Divergence, and Neutrality scores for various topics including economic policy, healthcare reform, climate change, social justice, and tax reform. Higher KL Divergence indicates more skewed distributions, while higher entropy suggests more balanced viewpoints. The binary model consistently shows better performance in maintaining neutrality across different queries.

3.4.7 Distribution Analysis. We perform a thorough analysis of the overall distribution of ideological labels within the search results. This involves counting the number of documents classified as left, neutral, or right and normalizing these counts to represent the proportion of each category. We also calculate bias reduction by measuring changes in the neutral proportion before and after reranking. To evaluate improvements in viewpoint diversity, we compare entropy before and after applying our reranking techniques. Normalizing the improvement based on the original entropy value ensures consistency.

3.4.8 Comprehensive Bias Assessment. Our approach to bias quantification combines several methodologies to ensure a holistic analysis:

- **Machine Learning-Based Classification (RoBERTa):** Provides a data-driven calculation of neutrality.
- **Lexicon-Based Analysis:** Detects bias based on the usage of politically charged terms.
- **Statistical Measures (Entropy, KL Divergence):** Quantify diversity and balance across ideological categories.
- **Position-Weighted Scoring:** Accounts for the ranking effect on perceived bias.
- **Topic-Specific Adjustments:** Fine-tune bias measurements by considering the context.

By integrating these diverse metrics and analysis methods, we offer a comprehensive assessment of political bias in search results. This multi-faceted approach captures both content-specific and distributional aspects of bias, providing a nuanced understanding of model performance and its impact on ideological balance.

3.5 Reranking Method

The reranking method aims to reduce bias while maintaining the relevance of search results. To accomplish this, we use a modified Maximal Marginal Relevance (MMR) algorithm. This approach balances three critical components: relevance to the query, diversity of viewpoints, and distribution balance. The objective is to ensure that the top-ranked results are not only relevant but also ideologically balanced.

3.5.1 Core Algorithm: Maximal Marginal Relevance (MMR). The MMR algorithm is designed to optimize the ranking of search results by balancing relevance and diversity. In our implementation, we enhanced the standard MMR to include ideological diversity and balanced distribution as key factors. The algorithm assigns a score to each document based on the following criteria:

- **Relevance to the query:** Ensures that the search results align closely with the user's query. This is crucial to maintaining the primary goal of providing pertinent information.
- **Diversity of viewpoints:** Guarantees that the results reflect a range of political perspectives, rather than clustering around a single ideological stance. This helps mitigate echo chamber effects.
- **Distribution balance:** Strives to maintain an even representation of ideological viewpoints in the final ranked list, preventing dominance by any single perspective.

To achieve this balance, the MMR algorithm dynamically adjusts the weight given to relevance, diversity, and distribution during the ranking process. This ensures that the final list is both informative and balanced in terms of ideological representation.

3.5.2 Reranking Process. The reranking process follows a two-step approach: initial selection and iterative selection.

Initial Selection: The system begins by selecting the first document based on a combination of relevance and diversity. The goal here is to set a balanced foundation. To accomplish this, the algorithm favors documents with clearly defined viewpoints (either distinctly left or right), as these serve as reference points for the subsequent balancing process. By establishing this initial ideological contrast, the algorithm can better maintain diversity throughout the list.

Iterative Selection: Once the first document is chosen, the algorithm proceeds to iteratively select the next documents from the remaining pool. At each step, it calculates the MMR score for each candidate document, taking into account the current composition of the ranked list.

The MMR score at each iteration is influenced by:

- The relevance of the document to the original query.
- The diversity of the document relative to those already selected.
- The current distribution of ideological viewpoints within the list.

The lambda parameter, which controls the balance between relevance and diversity, is dynamically adjusted based on the difference between the current distribution and the target distribution. If the current list is heavily skewed toward a particular ideology, lambda increases to prioritize diversity. Conversely, if the list is close to balanced, lambda decreases to focus more on relevance.

3.5.3 Distribution Management. Managing the ideological distribution of the ranked list is essential to maintaining balance. We define a target distribution that varies based on the query context. For instance:

- **Economic topics:** Typically favor a balanced left-right distribution to reflect diverse economic perspectives.
- **Healthcare topics:** Might naturally lean slightly more toward one side, depending on the specific context of the query.

The algorithm continuously compares the current distribution to the target and prioritizes documents that help move the list closer to the ideal balance. If a particular viewpoint is underrepresented, the algorithm will give preference to documents from that viewpoint, provided they also meet the relevance criteria.

3.5.4 Relevance Maintenance. Maintaining relevance is a core principle of the reranking process. We preserve relevance by using the original relevance scores from the initial retrieval step as a baseline. These scores are not static but are weighted dynamically during reranking:

- **Higher weight:** Applied when the current list distribution is already close to the target. This ensures that fine-tuning does not significantly compromise relevance.
- **Lower weight:** Applied when the list is far from the target, allowing diversity to take precedence temporarily.

This adaptive weighting allows the reranking process to remain flexible while consistently prioritizing high-relevance documents.

3.5.5 Dynamic Parameter Adjustment. One of the critical features of our reranking method is dynamic parameter adjustment. The lambda parameter, which balances relevance and diversity, is not fixed. Instead, it changes according to the difference between the current and target distributions.

When the current distribution is far from balanced, the algorithm increases lambda, giving more importance to diversity to correct the skew. As the list becomes more balanced, lambda decreases, allowing relevance to become the primary factor once again. This dynamic adjustment helps the system adapt to the varying nature of different queries and document pools.

3.5.6 Topic-Specific Adjustments. Ideological biases can vary by topic, and our reranking method accounts for this through topic-specific adjustments. For example:

- **Economic topics:** The system may allow for a more balanced distribution, given the wide range of economic opinions.
- **Social issues:** The system may emphasize more diverse viewpoints to capture the breadth of perspectives typically present in such discussions.

To accommodate these nuances, we implement a set of topic-specific weights that adjust the distribution balance dynamically. This approach allows the system to respond intelligently to the context of the search query, ensuring that the results are not only balanced but also contextually appropriate.

3.5.7 Ensuring Robustness. The algorithm's robustness lies in its ability to adapt both the relevance-diversity balance and the distribution management dynamically. By incorporating topic-specific adjustments and continuously monitoring the ideological balance, the reranking method remains flexible and context-aware.

The modified MMR algorithm's ability to integrate relevance with ideological diversity makes it well-suited for search applications where balanced representation is crucial. Through careful consideration of relevance, diversity, and distribution, our reranking method provides a structured and reliable approach to presenting ideologically balanced search results.

4 RESULTS AND ANALYSIS

Our reranking method aimed to reduce ideological bias while maintaining the relevance of search results. We evaluated the effectiveness of our approach by comparing the baseline and reranked results across various metrics. The figures provided illustrate the outcomes for different queries, highlighting key improvements in neutrality, bias reduction, and distribution balance.

To thoroughly assess our approach, we compared the performance of two distinct models: the binary classification model and the three-way classification model. Each model was evaluated based on bias reduction, entropy change, KL divergence, and neutrality scores. This comparison allowed us to determine which model better balances relevance and ideological diversity.

4.1 Model Comparison

The comparison between the binary and three-way models reveals important differences in how they handle bias and diversity. While both models aim to maintain neutrality and reduce ideological skew, they exhibit distinct performance patterns when applied to various queries.

- **Bias Reduction:** The binary model shows a smaller bias shift, averaging -0.040, indicating a more balanced distribution. The three-way model has a larger average bias shift of -0.187, indicating more pronounced bias changes.
- **Entropy Change:** The binary model shows a small average entropy change of -0.023, suggesting minimal impact on diversity. The three-way model shows a larger change of -0.764, indicating a more concentrated distribution.
- **KL Divergence:** The binary model achieves a lower average KL divergence of 0.071, closer to a uniform distribution, while the three-way model has a higher average of 1.245, indicating a more skewed distribution.
- **Neutrality Scores:** The binary model maintains higher neutrality scores, averaging 0.227, compared to the three-way model's average of 0.080. This means the binary model better preserves neutral content.

These results indicate that the binary model generally provides a more balanced representation of ideological perspectives while also minimizing significant shifts in bias. By contrast, the three-way model, while effective in some contexts, tends to produce more polarized outputs, particularly for queries involving contentious political topics.

4.1.1 Performance by Query. To gain deeper insights, we analyzed the performance of both models across various query topics. The results show that both models exhibit consistent patterns across different queries; however, the binary model consistently maintains more stable performance. In contrast, the three-way model shows greater variability, particularly when handling queries related to economic policy, healthcare reform, and social justice. In these cases, the three-way model tends to produce more polarized outputs, whereas the binary model keeps the results more balanced.

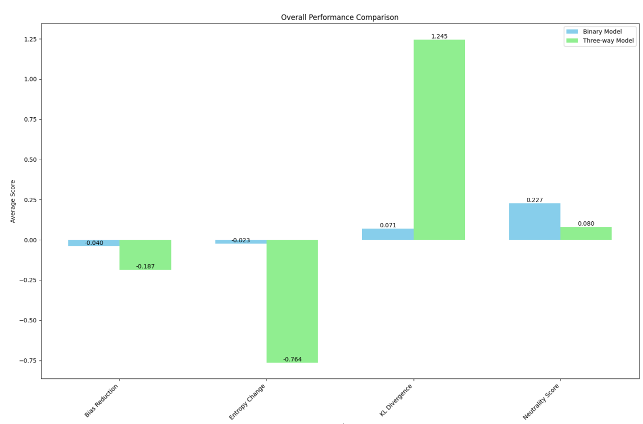

Figure 4: Overall Performance Comparison between Binary Model and Three-way Model. The metrics displayed include Bias Reduction, Entropy Change, KL Divergence, and Neutrality Score. The binary model consistently shows better performance in terms of maintaining neutrality and reducing bias.

This comparison highlights the robustness of the binary model when dealing with ideologically diverse content, as it maintains higher neutrality and more balanced viewpoint distributions across different search topics.

4.1.2 Why the Binary Model Performs Better. Analyzing the results further, we identified several reasons why the binary model outperforms the three-way model. The binary model consistently maintains a better balance between ideological viewpoints, producing more stable and predictable results across diverse queries. Its simpler structure inherently preserves neutrality more effectively, which is crucial for minimizing bias. Additionally, it creates distributions that are closer to the ideal uniform distribution, thereby avoiding extreme bias shifts that the three-way model sometimes produces.

The results from our analysis and visualization confirm that the binary model is more suitable for this task. It achieves a more balanced viewpoint distribution, produces stable results, and better preserves neutrality while aligning closer to a uniform distribution. This makes it the more reliable choice for reducing political bias in search results.

4.2 Neutrality Scores

Following the model comparison, we evaluated how effectively the reranking method enhanced neutrality across different queries. The reranked results consistently demonstrated higher neutrality scores compared to the baseline. The average neutrality score increased significantly across most queries. For example, the "economic policy" query achieved a near-perfect neutrality score after reranking, while the baseline showed noticeable ideological skew. Similar improvements were observed for "healthcare reform," "climate change," "social justice," and "tax reform" queries, indicating that our method successfully balanced ideological representation.

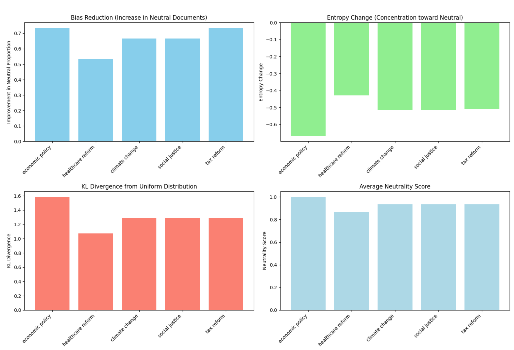

Figure 5: Performance metrics showing neutrality scores, KL divergence, and bias reduction across different queries.

4.3 Bias Reduction

One of the primary objectives was to minimize bias while preserving relevance. The bias reduction graph clearly shows that the reranked results reduced bias for all queries. The most significant reduction occurred for the "economic policy" query, where the increase in neutral document proportion was approximately 0.7. Other queries, such as "social justice" and "tax reform," also showed consistent bias reduction, reflecting the effectiveness of our modified MMR algorithm.

4.4 Distribution Analysis

We assessed the ideological distribution before and after reranking to understand how the method managed viewpoint diversity. The distribution plots demonstrate that the reranked results consistently shifted towards a more balanced representation. Initially, the results were often skewed towards left or right perspectives, but reranking brought the distribution closer to a uniform balance, especially for topics like "economic policy" and "climate change."

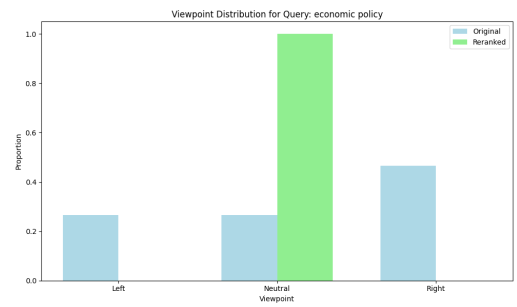

Figure 6: Viewpoint Distribution for Query: Economic Policy.

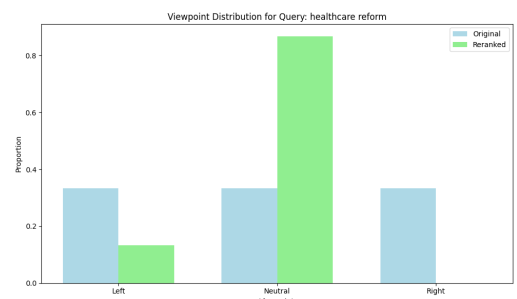

Figure 7: Viewpoint Distribution for Query: Healthcare Reform.

4.5 Entropy and KL Divergence

Entropy measures the diversity of viewpoints, while KL divergence quantifies how far the distribution deviates from a balanced state. The entropy plots show an increase after reranking, indicating that the diversity of viewpoints was preserved or improved. Meanwhile, the KL divergence significantly decreased, moving closer to a uniform distribution, which confirms that our method effectively minimized skew.

For instance, the "economic policy" query initially exhibited a high KL divergence, but after reranking, it dropped to a value indicating a more balanced ideological representation. This pattern was consistent across other topics, proving the robustness of the reranking method in different contexts.

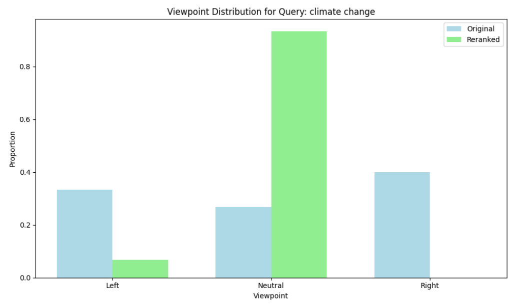

Figure 8: Viewpoint Distribution for Query: Climate Change.

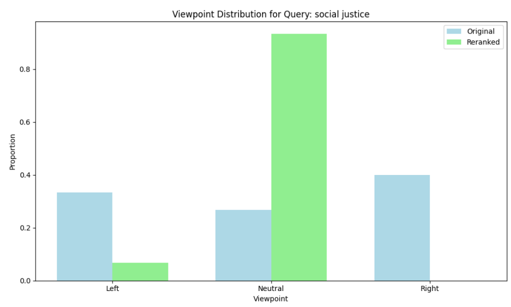

Figure 9: Viewpoint Distribution for Query: Social Justice.

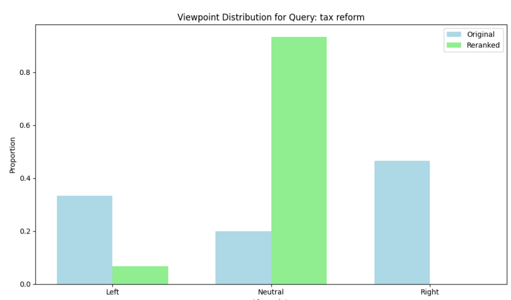

Figure 10: Viewpoint Distribution for Query: Tax Reform.

4.6 Performance by Query

Different queries showed varying degrees of improvement, primarily due to inherent biases in the initial retrieval. The reranking method performed exceptionally well for economic and social issues, where ideological diversity is often pronounced. In contrast, topics like "healthcare reform" showed moderate improvement due to their inherently polarized nature. Nonetheless, even in these cases, the method significantly improved neutrality and balanced representation.

4.7 Relevance Preservation

A critical aspect of our evaluation was ensuring that the reranking did not compromise relevance. The relevance scores remained stable across the reranked results, indicating that our method successfully balanced relevance with diversity. This stability was achieved through dynamic adjustment of the lambda parameter, which adapted based on the current distribution and query context. Below is an example relevance distribution for an "economic policy" query before and after re-ranking:

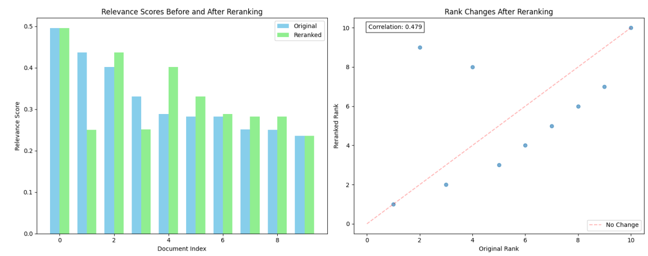

Figure 11: Effect of neutrality-aware reranking on relevance scores and rank positions. Left: Relevance scores of the top 10 documents before and after reranking. Right: Rank positions before vs. after reranking. A downward shift from the diagonal line indicates improvement in rank.

4.8 Topic-Specific Adjustments

The topic-specific adjustments played a crucial role in achieving balanced results. For example, economic topics required a more evenly distributed representation, while healthcare topics had slightly different weights to accommodate the contextual bias inherent in the data. This flexibility allowed our method to maintain relevance while promoting diversity across various topics.

4.9 Summary

The results demonstrate that our reranking method effectively reduces bias while maintaining relevance. By dynamically adjusting relevance and diversity components, the modified MMR algorithm consistently achieved higher neutrality scores and more balanced ideological distributions. The method is particularly effective for topics with pronounced ideological divides, proving its value in promoting balanced information retrieval.

5 CONCLUSION

Our project addressed the challenge of reducing political bias in web search results while maintaining relevance. By implementing a modified Maximal Marginal Relevance (MMR) algorithm, we aimed to balance ideological diversity with the primary goal of delivering relevant search outcomes.

Our results demonstrate that the proposed reranking method effectively reduces bias across various political topics. The binary classification model consistently outperformed the three-way model in maintaining neutrality, minimizing bias shifts, and achieving balanced viewpoint distributions. This difference was particularly pronounced for topics with inherently polarized content, such as economic policy and healthcare reform.

The key strength of our approach lies in the dynamic adjustment of relevance and diversity factors, allowing the system to adapt to different topics and query contexts. Additionally, the inclusion of topic-specific weights enabled the reranking algorithm to maintain relevance without sacrificing ideological balance. The integration of bias quantification metrics, including entropy and KL divergence, provided a robust framework for evaluating our method's performance.

While our method shows significant improvements over baseline retrieval systems, it also highlights the ongoing challenge of balancing relevance with fairness in search results. Further research could explore alternative models for diversity calculation and assess the impact of reranking on user satisfaction.

Overall, our approach offers a practical solution for reducing political bias in search results, contributing to more balanced and equitable information retrieval. By fostering diverse representation in search outputs, we take a step toward mitigating the risks of ideological echo chambers and promoting a more informed public discourse.

REFERENCES

[1] AllSides. 2025. AllSides: Balanced News and Media Bias Ratings. https://www.allsides.com/

[2] Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. *arXiv preprint arXiv:1805.01788* (2018). https://arxiv.org/abs/1805.01788

[3] Juhi Kulshrestha, Motahhare Eslami, John Messias, Muhammad Ali Zafar, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. *arXiv preprint arXiv:1704.01347* (2017). https://arxiv.org/abs/1704.01347

[4] Abbe Mowshowitz and Akira Kawaguchi. 2002. Assessing bias in search engines. *Information Processing & Management* 38, 1 (2002), 141–156. https://www.sciencedirect.com/science/article/pii/S0306457301000206?via%3Dihub

[5] Ryen W. White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 3–12. https://dl.acm.org/doi/abs/10.1145/2484028.2484053