Fair Summarization: Bridging Quality and Diversity in Extractive Summaries

Sina Bagheri Nezhad, Sayan Bandyapadhyay, Ameeta Agrawal
Department of Computer Science
Portland State University
Portland, Oregon, USA
{sina.bagherinezhad,sayanb,ameeta}@pdx.edu

Abstract

Fairness in multi-document summarization of user-generated content remains a critical challenge in natural language processing (NLP). Existing summarization methods often fail to ensure equitable representation across different social groups, leading to biased outputs. In this paper, we introduce two novel methods for fair extractive summarization: `FairExtract`, a clustering-based approach, and `FairGPT`, which leverages GPT-3.5-turbo with fairness constraints. We evaluate these methods using `Divsumm` summarization dataset of White-aligned, Hispanic, and African-American dialect tweets and compare them against relevant baselines. The results obtained using a comprehensive set of summarization quality metrics such as SUPERT, BLANC, SummaQA, BARTScore, and UniEval, as well as a fairness metric F, demonstrate that `FairExtract` and `FairGPT` achieve superior fairness while maintaining competitive summarization quality. Additionally, we introduce composite metrics (e.g., SUPERT+F, BLANC+F) that integrate quality and fairness into a single evaluation framework, offering a more nuanced understanding of the trade-offs between these objectives. This work highlights the importance of fairness in summarization and sets a benchmark for future research in fairness-aware NLP models.

1 Introduction

Multi-document summarization, which condenses multiple documents into a concise summary, is a fundamental task in natural language processing (NLP). Summarization methods are typically either *extractive*, selecting the most important sentences, or *abstractive*, where the content is rephrased.

Early research focused on summarizing formal text sources such as news articles. However, with the rise of social media, attention has shifted to summarizing user-generated content, which is diverse in style and language. Social media platforms bring together users from varied backgrounds, introducing linguistic diversity through informal language, slang, and emojis. This diversity raises the challenge of ensuring fairness in summarization ensuring balanced representation of various social groups. In social media, where public opinion is shaped, fair summaries are essential to include different perspectives and avoid underrepresentation of one or more social groups.

In the context of social media, where millions of users contribute diverse perspectives, ensuring representation of this diversity in summaries becomes crucial. Social media platforms encompass a wide range of voices, including those from historically underrepresented or marginalized groups, making it essential that summarization methods capture this diversity fairly. Without proper representation, certain voices might be excluded or misrepresented, leading to biased summaries that skew public discourse (Binns, 2017; Hutchinson and Mitchell, 2018). The need for fairness in

Algorithmic Fairness through the Lens of Metrics and Evaluation Workshop
@ The 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, Canada.

ChatGPT-EXT (Zhang et al., 2023)	FairGPT (Ours)
If you see on the news something about the Chicago Kitchen Clown Bandits then it will be referring me my friend Eten and I. Turns out not all White Castles are the same. Why do you push me away Chicago?! I mean I'm from Chicago. I'll cheer for the Bears, but I'm a bigger 49ers fan. Is this new wave of Chicago Rap gonna be like the Hyphy movement? Don't talk shot about Chicago, or those big shoulders will plow right into your little Boston ass. Nothing makes me happier than seeing the Bulls win #ChicagoBasketball #Bullieve.	Don't talk shot about Chicago, or those big shoulders will plow right into your little Boston ass. Nothing makes me happier than seeing the Bulls win #ChicagoBasketball #Bullieve. Truuu we tryna find sum to do too.. I dnt wanna b n Chicago if ain't nobody here. Turns out not all White Castles are the same. Why do you push me away Chicago?! I mean I'm from Chicago. I'll cheer for the Bears, but I'm a bigger 49ers fan. Is this new wave of Chicago Rap gonna be like the Hyphy movement?

Table 1: Comparison of summaries by ChatGPT-EXT and FairGPT. Tweets from different groups are highlighted: Group 1 (e.g., White-aligned) and Group 2 (e.g., African-American).

summarization is further heightened by the fact that user-generated content is often informal and marked by dialectal variations, requiring models to go beyond traditional summarization approaches (Pitsilis et al., 2018). Therefore, ensuring that all groups—across race, gender, and linguistic diversity—are fairly represented is critical for generating balanced summaries that reflect the diversity of public opinion (Dash et al., 2018).

Despite advancements, bias remains a concern in automated summarization (Dash et al., 2019; Jung et al., 2019; Keswani and Celis, 2021; Olabisi et al., 2022) as most existing summarization methods focus on quality but fall short in optimizing fairness. This gap leads to the key question: if a summarization method is optimized for fairness, how does it affect the overall summary quality?

Previous studies suggest a trade-off between fairness and quality (Jung et al., 2019). Improving fairness can sometimes lower quality. While existing algorithms have made strides in balancing the two, none achieve perfect fairness.

In this paper, we address two research questions:

1. How does achieving perfectly fair summaries affect overall quality?
2. How well do current methods perform when considering both fairness and quality?

To illustrate the performance of fairness-aware summarization models, we compare summaries generated by ChatGPT-EXT (Zhang et al., 2023) and our proposed FairGPT model on a sample instance from `Divsumm` dataset (Olabisi et al., 2022). As shown in Table 1, FairGPT ensures equal representation of tweets from different groups, while ChatGPT-EXT shows a slight imbalance.

We make the following contributions:

- We propose `FairExtract`, a fair clustering-based extractive summarization method that achieves perfect fairness and is evaluated against baseline models using standard and composite quality-fairness metrics.
- We develop `FairGPT`, a large language model-based extractive summarization method that enforces fairness through equal representation and accurate content extraction using the longest common subsequence.
- We introduce composite metrics combining normalized quality scores with fairness, providing a comprehensive analysis of the quality-fairness trade-off in summarization models.

2 Related Work

The field of natural language processing (NLP) has increasingly focused on addressing bias and fairness, driven by the demand for equity in AI systems. Research has explored two key dimensions: intrinsic bias, stemming from text representations, and extrinsic bias, reflecting performance disparities across demographic groups (Han et al., 2023).

Early work on fairness in summarization (Shandilya et al., 2018; Dash et al., 2019) revealed that summaries often fail to represent source data fairly, even when source texts from different groups have similar quality. This led to the development of fairness-aware algorithms across various stages of summarization, including pre-processing, in-processing, and post-processing techniques. For example, Keswani and Celis (2021) proposed a post-processing method to mitigate dialect-based biases. Olabisi et al. (2022) introduced the DivSumm dataset, focusing on dialect diversity in summarization and evaluating algorithms on fairness.

Recent work has explored bias related to the position of input data. Olabisi and Agrawal (2024) studied position bias in multi-document summarization, showing that the order of input texts affects fairness. Similarly, Huang et al. (2023) analyzed clustering-based summarization models, which may introduce political or opinion bias, emphasizing the need for fair representation.

Fair clustering, another key technique, has also seen significant research. Chierichetti et al. (2017) introduced the concept of fairlets—small, balanced clusters that ensure fair representation across protected groups. Building on this, Chen et al. (2019) proposed proportional centroid clustering to eliminate biases in cluster-based models.

Further advancements include scalable techniques for fair clustering, such as the fair k-median clustering method (Backurs et al., 2019), and approaches that generalize fairness constraints across multiple protected groups (Bera et al., 2019). Esmaeili et al. (2020) extended this work to probabilistic fair clustering, offering solutions for uncertain group memberships.

In the domain of clustering methodologies, Micha and Shah (2020) explored fairness in centroid clustering, while Li et al. (2020) proposed Deep Fair Clustering (DFC), which leverages deep learning to filter sensitive attributes, improving both fairness and performance. This underscores the growing importance of combining fairness with robust clustering methods in NLP tasks.

3 Task Formulation

In this work, we address the challenge of diversity-preserving multi-document extractive summarization. Given a collection of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ from two diverse social groups, G_1 and G_2, the goal is to produce an extractive summary $\mathcal{S} = \{s_1, s_2, \ldots, s_k\} \subset \mathcal{D}$ of length $k << n$, ensuring balanced representation from both groups.

In this context, each document is a tweet from a specific dialect group, which serves as an indicator of its social group. Traditionally, various metrics like ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019) have been used to evaluate summary quality. However, our primary focus is on balancing both quality and fairness, particularly in terms of representing different social groups equitably. To measure fairness, we use the *Representation Gap (RG)* metric, as proposed by Olabisi et al. (2022). This metric captures how well the summary reflects the proportions of the original groups. A lower RG score indicates better balance and thus a fairer summary.

For a summary $\mathcal{S}$ of length k, let $N_1(\mathcal{S})$ and $N_2(\mathcal{S})$ represent the number of documents from groups G_1 and G_2, respectively. The Representation Gap is defined as:

$$\mathrm{RG}(\mathcal{S}) = \frac{\max\{N_1(\mathcal{S}), N_2(\mathcal{S})\} - \min\{N_1(\mathcal{S}), N_2(\mathcal{S})\}}{k}. \tag{1}$$

For example, if $k = 6$, with 4 documents from G_1 and 2 from G_2, the RG is 0.333. When both groups are equally represented, the RG is 0, indicating a *perfectly fair* summary.

Our analysis faces two key challenges: (1) While quality metrics improve with larger values, fairness improves with smaller Representation Gap (RG) values. (2) Quality and fairness metrics differ greatly in scale, making direct comparison difficult.

To address these issues, we introduce a new fairness metric, F, defined as:

$$F(\mathcal{S}) = 1 - \mathrm{RG}(\mathcal{S}) \tag{2}$$

This transformation ensures that larger F values indicate better fairness, aligning it with the behavior of quality metrics. Furthermore, we apply min-max normalization to rescale all metrics to the range $[0, 1]$, ensuring comparability across different scales. The normalization formula is given by:

$$\frac{\text{value} - \min}{\max - \min} \tag{3}$$

where $\min$ and $\max$ are the minimum and maximum observed values for the respective metric.

Finally, we introduce composite metrics, such as **SUPERT+F**, **BLANC+F**, **SummaQA+F**, **BARTScore+F**, and **UniEval+F**, which are the averages of the normalized quality metrics (e.g., SUPERT (Gao et al., 2020), BLANC (Vasilyev et al., 2020), SummaQA (Scialom et al., 2019), BARTScore (Yuan et al., 2021), and UniEval (Zhong et al., 2022)) and the fairness score F, providing a balanced assessment of both quality and fairness.

4 Fair Extractive Summarizers

In this work, we introduce two novel methods for fair extractive summarization: FairExtract and FairGPT. FairExtract utilizes clustering techniques with fairlet decomposition to ensure diversity in summaries while maintaining high-quality representation across different groups. FairGPT, on the other hand, leverages large language models (LLMs) such as GPT-3.5, incorporating fairness constraints and the longest common subsequence (LCS) method to match and fairly select content from different groups. Both methods prioritize fairness and ensure equitable representation in the generated summaries. The implementation code for both methods is available. [1]

4.1 FairExtract: A Clustering-based Fair Extractive Summarization Method

The task of clustering is central to the FairExtract process, which aims to generate diversity-preserving summaries. The method combines document embeddings, fairlet decomposition, and clustering techniques to ensure both fairness and quality. Below, we describe the steps involved in detail:

1. **Embedding Documents:** We begin by embedding each document (tweet) into a high-dimensional space (e.g., using a pretrained model such as BERT (Devlin et al., 2019)), capturing its semantic content in Euclidean space. This embedding enables us to compute meaningful distances between documents, which is crucial for clustering.
2. **Fairlet Decomposition:** To ensure fairness in the summarization process, we decompose the dataset into fairlets. A fairlet is the smallest set of documents that maintains proportional balance between two groups, G_1 and G_2 (Backurs et al., 2019). If the proportions of G_1 and G_2 are g_1 and g_2, respectively, where $\gcd(g_1, g_2) = 1$, a fairlet must contain exactly g_1 documents from G_1 and g_2 documents from G_2. This ensures that the composition of the fairlet reflects the required ratio between the two groups, maintaining fairness at the smallest possible scale. The decomposition aims to minimize the sum of Euclidean distances between documents within the same fairlet.
3. **Finding the Fairlet Center:** Once the dataset is divided into fairlets, we compute the center of each fairlet. The center is the document within the fairlet that minimizes the sum of distances to all other documents in the same fairlet. This document acts as the representative of the fairlet, summarizing the content while maintaining group balance.
4. **k-Median Clustering on Fairlet Centers:** After identifying the centers of all fairlets, we apply the k-median clustering algorithm to these centers. In the k-median problem, we are given a set of points P in a d-dimensional space, and we aim to partition them into k clusters $\Pi = \{P_1, \ldots, P_k\}$ that minimize the following cost:

$$\min_{C \subset P : |C| = k} \sum_{c_i \in C | 1 \leq i \leq k} \sum_{p \in P_i} ||p - c_i||. \tag{4}$$

[1] https://github.com/PortNLP/FairEXTSummarizer

The number of clusters k is selected such that $k \times (g_1 + g_2)$ equals the desired number of documents in the summary. This step ensures that the clusters formed are representative of both social groups.

5. **Summary Construction:** From each k-median cluster, we select the center fairlet and include all documents within that fairlet in the final summary. By selecting one fairlet from each cluster, we maintain both quality and fairness, as the summary reflects the balanced representation of both groups. The resulting extractive summary ensures that the most salient information is captured while maintaining equitable representation of the social groups.

For a formal representation of the process, see Algorithm 2 in Appendix A.1.

4.2 FairGPT: An LLM-based Fair Extractive Summarization Method

FairGPT leverages GPT-3.5-turbo to generate fair extractive summaries by selecting an equal number of sentences from different social groups. It applies fairness checks and uses the longest common subsequence (LCS) to match generated summaries with the original tweets. Below are the detailed steps:

1. **Input Preparation:** The dataset is split into two groups (e.g., White-aligned and Hispanic dialects), and a document with sentences for each group is created as input for the summarization process.
2. **Summarization using an LLM:** We use an LLM (GPT-3.5-turbo) to generate a summary of length L, selecting $L/2$ sentences from each group to ensure balanced representation. The specific prompt used for this task is available in the Prompt 1.
3. **Matching using Longest Common Subsequence (LCS):** As GPT sometimes generates partial sentences, we apply LCS to match the generated summary with the closest original tweets. The full tweets corresponding to the longest common subsequences are added to the final summary.
4. **Output Check:** After generating the summary, we verify two key aspects. First, at least 50% of the content in each GPT-generated sentence must match the corresponding original tweet using the LCS. Second, we ensure that the summary is perfectly fair, with equal representation from each group.

   This output check is crucial because large language models, such as GPT-3.5-turbo, sometimes generate unexpected outputs that do not align with the input instructions. To ensure the generated summaries meet both fairness and content similarity criteria, we repeat the process if either condition is not satisfied. In our tests of generating 75 summaries, the repetition process never exceeded 10 iterations, and the average number of repetitions across all tests was 1.6, indicating the efficiency and reliability of the output check mechanism.
5. **Final Output:** Once the summary satisfies both fairness and similarity requirements, it is saved as the final output.

For a formal representation of the process, see Algorithm 1.

FairGPT Prompt

```
system:   "You are an extractive fair summarizer that follows the
          output pattern. A fair summarizer should select the same
          number of sentences from each group of people."

user:     "Please extract sentences as the summary.
          The summary should contain {L} sentences which means
          select {L/2} number of sentences from each group of people
          to represent the idea of all groups in a fair manner.
          Document:{document}"
```

Prompt 1: Prompt used in FairGPT. The variable L refers to the total number of sentences to be extracted.

```
Algorithm 1 FairGPT Algorithm
Input:
    • Document set $\mathcal{D}$ divided into groups $G_1$ and $G_2$
    • Desired summary length $L$ with $L/2$ sentences from each group
Output: Fair extractive summary $\mathcal{S}$
Step 1: Input Preparation
Create documents for $G_1$ and $G_2$, clearly labeling each sentence based on its group.
Step 2: Summarization using LLM
Instruct LLM (GPT-3.5-turbo) using Prompt 1 to select $L/2$ sentences from each group, ensuring fair representation.
Step 3: Matching using Longest Common Subsequence (LCS)
Use LCS to match the GPT-generated sentences with the original dataset to identify the closest matching tweets and include the full sentences in the summary.
Step 4: Ensuring 50% Similarity
Ensure that at least 50% of the content in each generated sentence matches the corresponding original tweet using LCS.
Step 5: Fairness Check
Verify that the summary contains an equal number of sentences from $G_1$ and $G_2$. If fairness or similarity conditions are not met, go to Step 2.
Step 6: Final Output
Save the final summary $\mathcal{S}$ once both fairness and quality thresholds are satisfied.
Return: The final summary $\mathcal{S}$.
```

5 Experimental Setup

Next, we describe the dataset, baseline methods, and evaluation metrics that are used to comprehensively assess the quality and fairness of the generated summaries.

5.1 Dataset

The dataset used in this study is *DivSumm* (Olabisi et al., 2022), consisting of tweets from three ethnic groups—White-aligned, Hispanic, and African-American—across 25 topics, with 30 tweets per group per topic, totaling 2,250 tweets. This diversity allows us to evaluate the fairness of our summarization methods across varied social and cultural contexts.

Our model works with two groups at a time, so we explore three pairings: White-Hispanic, Hispanic-African American, and White-African American. Each pairing maintains proportional representation from both groups to ensure an equitable balance in the summarization process.

For our experiments, we formed 60 tweets per group pair (30 from each group) and generated a 6-tweet summary per pair, covering all 25 topics. This yielded 75 distinct summaries per model, allowing us to evaluate both fairness and quality comprehensively. A sample of the dataset is available in Appendix A.2.

5.2 Baseline Methods

Here, we provide a detailed description of the baseline methods used in our comparative analysis:

Naive: In the Naive baseline method, L tweets are randomly chosen from the input without any specific criteria. This approach represents a straightforward, non-strategic selection process and serves as a basic reference point for evaluating other methods.

NaiveFair: The NaiveFair baseline method involves randomly selecting $L/2$ tweets from each social group. This method ensures equal representation from each group, providing a basic notion of fairness without any sophisticated processing.

For the Naive and NaiveFair methods, which involve randomness in selecting summaries, we conducted the experiment five times for each summary, resulting in 375 different summaries for each of these methods.

TextRank: TextRank is an unsupervised graph-based ranking method used for extractive summarization (Mihalcea and Tarau, 2004). This standard `vanilla` baseline approach uses a single ag-

gregated set of randomized documents from all groups as input for summarization, without any pre-processing.
BERT-Ext: BERT-Ext is an extractive summarization model that utilizes pre-trained embeddings from BERT and k-means clustering to select sentences closest to the centroid as summaries (Miller, 2019). Similar to the TextRank baseline, we implemented BERT-Ext `vanilla` method.
Cluster-Heuristic (Cluster-H): This method first partitions the input documents into group-based subsets before generating separate group summaries of length . These group-level summaries are shuffled, combined and then used to generate a final, unified summary (Dash et al., 2019; Olabisi et al., 2022). As summarization models, we use TextRank and BERT-Ext.
Cluster-Automatic (Cluster-A): In this attribute-agnostic approach, documents are clustered automatically into m subsets, and corresponding summaries of length are generated. The summaries are concatenated and used to generate a final summary (Olabisi et al., 2022). As summarization models, we experiment with TextRank and BERT-Ext.
ChatGPT-EXT: This approach uses GPT-3.5 for extractive summarization by employing in-context learning and chain-of-thought reasoning to identify key sentences. It focuses on extracting salient content from documents to generate coherent summaries while maintaining the structure of the original text (Zhang et al., 2023).

5.3 Evaluation Metrics

Below, we list the several reference-free metrics which do not rely on human-written reference text used for evaluation in this study.

- **SUPERT:** SUPERT (Gao et al., 2020) evaluates the quality of a summary by measuring its semantic similarity with a pseudo reference summary. It employs contextualized embeddings and soft token alignment techniques, providing an in-depth analysis of the semantic fidelity of generated summaries.
- **BLANC:** BLANC (Vasilyev et al., 2020) is a reference-less metric that measures the improvement in a pretrained language model's performance during language understanding tasks when given access to a summary.
- **SummaQA:** SummaQA (Scialom et al., 2019) employs a question-answering model based on BERT to answer cloze-style questions using the system-generated summaries, providing insights into the summarization's factual accuracy and coherence.
- **BARTScore:** BARTScore (Yuan et al., 2021) is a parameter- and data-efficient metric that supports the evaluation of generated text from multiple perspectives, including informativeness and coherence.
- **UniEval:** UniEval (Zhong et al., 2022) is a unified multi-dimensional evaluator that reframes natural language generation evaluation as a Boolean Question Answering (QA) task, guiding the model with different questions to evaluate from multiple dimensions. It is reference-free in three dimensions (coherence, consistency, fluency), but not relevance. For our evaluation, we focused on the reference-free dimensions of UniEval and reported the overall average performance.
- **Fairness (F):** To align fairness with the quality metrics, we define $F = 1 - \text{RG}$, where larger values represent better fairness. The Representation Gap (RG) metric (Olabisi et al., 2022) assesses the fairness of summaries by measuring the balance in the representation of different groups.
- **Composite Metrics (Metric+F):** For each quality metric (e.g., SUPERT, BLANC, SummaQA, BARTScore, and UniEval), we introduce a composite metric that combines the normalized quality score with the fairness score F. These composite metrics, such as **SUPERT+F**, **BLANC+F**, **SummaQA+F**, **BARTScore+F**, and **UniEval+F**, are computed by taking the average of the normalized quality metric and the fairness score F. A higher value of these composite metrics reflects a better balance between the summary's quality (as measured by the respective metric) and fairness.

6 Results and Discussion

In this section, we present the results of our evaluation, comparing the performance of various summarization models on both quality and fairness metrics.

Model	SUPERT	BLANC	SummaQA	BARTScore	UniEval	F
Naive	0.525	0.135	0.063	-1.788	0.391	0.732
NaiveFair	0.526	0.137	0.065	-1.776	0.386	**1.000**
TextRank Vanilla	0.527	0.108	**0.081**	-1.852	0.401	0.727
TextRank Cluster-A	0.530	0.107	0.075	-1.827	0.383	0.693
TextRank Cluster-H	0.530	0.107	0.077	-1.922	0.387	0.709
BERT-EXT Vanilla	0.544	0.137	0.070	-1.427	0.396	0.680
BERT-EXT Cluster-A	0.553	0.138	0.071	-1.535	0.399	0.728
BERT-EXT Cluster-H	0.554	0.133	0.070	-1.486	0.365	0.689
ChatGPT-EXT	**0.668**	**0.140**	0.065	**-0.642**	**0.434**	0.698
FairExtract (Ours)	0.530	**0.140**	0.066	-1.801	0.411	**1.000**
FairGPT (Ours)	0.644	0.139	0.075	-0.821	0.418	**1.000**

Table 2: Evaluation results for various summarization methods. The best values for each metric are shown in bold.

6.1 Results of Quality and Fairness

The models were assessed using SUPERT, BLANC, SummaQA, BARTScore, UniEval, and the fairness metric F. Table 2 presents the results.

Naive and NaiveFair Baselines: The Naive baseline, which randomly selects sentences without any fairness consideration, performs relatively poorly across most quality metrics, particularly on SummaQA and BARTScore, where it scores significantly lower. However, it achieves a reasonable fairness score ($F = 0.732$), despite its lack of sophisticated fairness mechanisms. The NaiveFair model, which ensures equal representation from both groups, shows a slight improvement in fairness, achieving the maximum F value of 1. However, this fairness comes at a slight cost to quality, as it falls behind on some metrics like UniEval.

TextRank Models: The TextRank Vanilla method shows a balanced performance in terms of quality, with the highest SummaQA score (0.081), but suffers in BLANC and BARTScore. Variations of TextRank, such as Cluster-A and Cluster-H, show slight improvements in specific metrics like SUPERT and BLANC, but they still struggle in ensuring fairness, with scores in the range of $F = 0.693$ to $F = 0.727$.

BERT-Ext Models: The BERT-EXT models generally outperform the TextRank methods in quality metrics. BERT-EXT Vanilla achieves higher SUPERT and BARTScore scores compared to TextRank, with BERT-EXT Cluster-A further improving on these metrics, particularly in SUPERT (0.553) and BLANC (0.138). However, the fairness scores for these models remain moderate, with F values ranging from 0.680 to 0.728, indicating room for improvement in terms of group representation balance.

ChatGPT-Ext: The ChatGPT-Ext method stands out as the top performer in terms of quality, achieving the highest scores in SUPERT (0.668), BLANC (0.140), BARTScore (-0.642), and UniEval (0.434). This demonstrates its effectiveness in producing semantically rich and coherent summaries. However, its fairness score of $F = 0.698$ indicates that while it excels in quality, there is still room for improvement in terms of group representation.

FairExtract and FairGPT (Ours): Our proposed models, FairExtract and FairGPT, were designed with fairness as a core objective. Both models achieve perfect fairness, with $F = 1$, while still maintaining competitive quality. FairExtract performs comparably to TextRank in terms of quality metrics, excelling in BLANC (0.140) and achieving respectable scores in SUPERT and UniEval. FairGPT, leveraging the power of GPT-3.5, shows a strong balance between quality and fairness, with particularly high SUPERT (0.644) and BARTScore (-0.821) scores. These results suggest that our models successfully balance the trade-off between quality and fairness, making them robust options for fairness-aware summarization tasks.

Overall, ChatGPT-Ext achieves the highest quality metrics, while FairExtract and FairGPT lead in fairness without compromising quality; notably, FairGPT emerges as the best model, striking an optimal balance between quality and diversity, underscoring the success of our proposed methods in achieving fair and high-quality summarizations.

6.2 Results Aggregating Quality and Fairness

The composite evaluation metrics are presented in Table 3. These metrics aggregate both quality and fairness, both receiving equal weight (50%) in the overall score. Our results show that FairExtract,

Clustering-based Methods					
Model	SUPERT+F	BLANC+F	SumQA+F	BARTSc+F	UniEval+F
Naive	0.585	0.609	0.468	0.713	0.601
NaiveFair	0.720	0.749	0.606	**0.848**	0.732
TextRank Vanilla	0.585	0.531	0.494	0.703	0.605
TextRank Cluster-A	0.571	0.513	0.467	0.689	0.577
TextRank Cluster-H	0.579	0.521	0.478	0.687	0.588
BERT-EXT Vanilla	0.582	0.590	0.453	0.725	0.578
BERT-EXT Cluster-A	0.616	0.615	0.479	0.737	0.604
BERT-EXT Cluster-H	0.598	0.583	0.457	0.723	0.564
FairExtract (Ours)	**0.724**	**0.758**	**0.607**	0.845	**0.747**
LLM-based Methods					
ChatGPT-EXT	0.737	0.607	0.454	0.817	0.611
FairGPT (Ours)	**0.837**	**0.760**	**0.615**	**0.945**	**0.751**

Table 3: Evaluation results using composite metrics for clustering-based and LLM-based summarization methods with equal weighting of quality and fairness ($\alpha = 0.5$). The best values for each metric are highlighted in bold.

the proposed clustering-based summarization method, consistently outperforms other clustering-based models across most composite metrics, including SUPERT+F, BLANC+F, SummaQA+F, and UniEval+F. Although NaiveFair scores slightly higher on BARTScore+F, the difference is minimal, at just 0.003 (or 0.35% in percentage terms), indicating that FairExtract achieves near-optimal performance in balancing quality and fairness.

Similarly, among the large language model (LLM)-based methods, FairGPT stands out as the best performer, achieving the highest composite scores across almost all metrics, including SUPERT+F, BLANC+F, SummaQA+F, BARTScore+F, and UniEval+F. This demonstrates that FairGPT effectively balances quality and fairness, setting a new benchmark in fair summarization using LLMs.

To assess the impact of varying the weight on fairness, we explored a composite metric formula: $(1-\alpha) \times \text{Quality} + \alpha \times F$, where α controls the fairness weight. When $\alpha = 0.5$, fairness and quality are equally weighted, as in the results presented in Table 3. We further experimented with reducing the fairness weight to find the minimum value of α at which FairExtract still outperforms other clustering-based methods.

Table 5 in Appendix A.3 shows the results for $\alpha = 0.16$ (i.e., a 16% fairness incentive). Even with this reduced fairness weight, FairExtract continues to outperform all clustering-based methods across most metrics. Similarly, FairGPT remains the best-performing LLM-based method, maintaining dominance even with the lower fairness incentive.

In summary, our experimental results clearly demonstrate that FairExtract and FairGPT, the two fair summarization models proposed in this paper, achieve a robust balance between quality and fairness across multiple metrics. FairExtract consistently surpasses other clustering-based models when fairness is weighted equally with quality, while FairGPT sets new benchmarks among LLM-based methods, showing superior performance in both quality and fairness. Even when the fairness incentive is reduced to 16%, FairExtract continues to perform better than most competing models, underscoring the strength of our approach in ensuring diverse representation without compromising summary quality. These findings highlight the importance of incorporating fairness into summarization tasks and demonstrate the effectiveness of our proposed methods in achieving this balance.

7 Conclusion

In this paper, we introduced two novel methods, FairExtract and FairGPT, to address the critical challenge of fairness in multi-document extractive summarization. Both methods were designed to ensure equitable representation of social groups while maintaining competitive summarization quality. Our extensive experiments demonstrated that both FairExtract and FairGPT achieve perfect fairness without significantly compromising on standard quality metrics.

We also introduced new composite metrics (e.g., SUPERT+F, BLANC+F) that combine quality and fairness scores, offering a more nuanced evaluation of the trade-offs between these two dimensions. The results showed that our methods strike a strong balance between quality and fairness, with

FairExtract performing exceptionally well in clustering-based approaches and FairGPT setting new benchmarks among LLM-based methods.

These findings highlight the importance and feasibility of integrating fairness into summarization tasks, where diverse representation is crucial. Future work can build on these models by extending them to abstractive summarization, exploring additional fairness constraints, and applying them to larger, more diverse datasets. Our work serves as a significant step toward building fair and inclusive summarization systems for real-world applications.

8 Limitations

While FairExtract and FairGPT show advances in ensuring fairness in multi-document summarization, several limitations remain.

First, our methods focus on extractive summarization, which, while preserving input fidelity, may not capture the semantic richness of abstractive methods (Lebanoff et al., 2019). Extending our approach to abstractive models presents additional challenges, particularly in balancing fairness with coherence and fluency.

Second, the dataset consists of social media content, which may limit generalization to other domains like news or scientific articles. The informal nature of social media language introduces variability that might not translate to more formal text types.

Third, our work focuses on monolingual inputs, specifically in English. Future research could extend these methods to multilingual inputs, where additional factors such as language diversity and cross-lingual transfer, as highlighted by Bagheri Nezhad and Agrawal (2024), would need to be addressed to ensure fairness across languages.

Additionally, while we employ standard quality and fairness metrics, they do not fully capture subjective factors such as readability or user trust. Human evaluation could provide deeper insights into the practical implications of fairness and quality.

Finally, the computational complexity of fair clustering and large language models may limit scalability in real-time or resource-constrained environments.

Despite these challenges, our work marks a significant step toward fairer summarization models, and addressing these limitations could enhance the robustness of fairness in NLP.

Acknowledgments and Disclosure of Funding

We would like to thank Aravind Inbasekaran for his valuable assistance throughout this project. This research has been supported by the National Science Foundation under Grant No. AF 2311397 and CRII:RI 2246174.

References

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. (2019). Scalable fair clustering. In *International Conference on Machine Learning*, pages 405–413. PMLR.

Bagheri Nezhad, S. and Agrawal, A. (2024). What drives performance in multilingual language models? In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Zampieri, M., Nakov, P., and Tiedemann, J., editors, *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 16–27, Mexico City, Mexico. Association for Computational Linguistics.

Bera, S., Chakrabarty, D., Flores, N., and Negahbani, M. (2019). Fair algorithms for clustering. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Binns, R. (2017). Fairness in machine learning: Lessons from political philosophy. *Decision-Making in Computational Design & Technology eJournal*.

Chen, X., Fain, B., Lyu, L., and Munagala, K. (2019). Proportionally fair clustering. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1032–1041. PMLR.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. (2017). Fair clustering through fairlets. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S., and Chakraborty, A. (2018). Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3:1 – 28.

Dash, A., Shandilya, A., Biswas, A., Ghosh, K., Ghosh, S., and Chakraborty, A. (2019). Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Esmaeili, S., Brubach, B., Tsepenekas, L., and Dickerson, J. (2020). Probabilistic fair clustering. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12743–12755. Curran Associates, Inc.

Gao, Y., Zhao, W., and Eger, S. (2020). SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Han, X., Baldwin, T., and Cohn, T. (2023). Fair enough: Standardizing evaluation and model selection for fairness research in NLP. In Vlachos, A. and Augenstein, I., editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–312, Dubrovnik, Croatia. Association for Computational Linguistics.

Huang, N., Tian, L., Fayek, H., and Zhang, X. (2023). Examining bias in opinion summarisation through the perspective of opinion diversity. In Barnes, J., De Clercq, O., and Klinger, R., editors, *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Hutchinson, B. and Mitchell, M. (2018). 50 years of test (un)fairness: Lessons for machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Jung, T., Kang, D., Mentch, L., and Hovy, E. (2019). Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3324–3335, Hong Kong, China. Association for Computational Linguistics.

Keswani, V. and Celis, L. E. (2021). Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 3802–3814, New York, NY, USA. Association for Computing Machinery.

Lebanoff, L., Song, K., Dernoncourt, F., Kim, D. S., Kim, S., Chang, W., and Liu, F. (2019). Scoring sentence singletons and pairs for abstractive summarization. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy. Association for Computational Linguistics.

Li, P., Zhao, H., and Liu, H. (2020). Deep fair clustering for visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Micha, E. and Shah, N. (2020). Proportionally Fair Clustering Revisited. In Czumaj, A., Dawar, A., and Merelli, E., editors, *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, volume 168 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 85:1–85:16, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into text. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Miller, D. (2019). Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Olabisi, O. and Agrawal, A. (2024). Understanding position bias effects on fairness in social multi-document summarization. In Scherrer, Y., Jauhiainen, T., Ljubešić, N., Zampieri, M., Nakov, P., and Tiedemann, J., editors, *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 117–129, Mexico City, Mexico. Association for Computational Linguistics.

Olabisi, O., Hudson, A., Jetter, A., and Agrawal, A. (2022). Analyzing the dialect diversity in multi-document summaries. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., and Na, S.-H., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018). Effective hate-speech detection in twitter data using recurrent neural networks. *Applied Intelligence*, 48:4730 – 4742.

Scialom, T., Lamprier, S., Piwowarski, B., and Staiano, J. (2019). Answers unite! unsupervised metrics for reinforced summarization models. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

Shandilya, A., Ghosh, K., and Ghosh, S. (2018). Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 97–98, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Vasilyev, O., Dharnidharka, V., and Bohannon, J. (2020). Fill in the BLANC: Human-free quality estimation of document summaries. In Eger, S., Gao, Y., Peyrard, M., Zhao, W., and Hovy, E., editors, *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20, Online. Association for Computational Linguistics.

Yuan, W., Neubig, G., and Liu, P. (2021). Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Zhang, H., Liu, X., and Zhang, J. (2023). Extractive summarization via ChatGPT for faithful summary generation. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., Zhu, C., Ji, H., and Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix / supplemental material

A.1 Fair Extract Formal Algorithmic Processes

In this section, we provide a detailed breakdown of the formal procedures used in our proposed method, FairExtract. These algorithm ensure fairness and quality in extractive summarization, addressing the core objectives of balanced representation and high-quality content extraction from diverse groups.

The FairExtract algorithm utilizes clustering techniques combined with fairlet decomposition to ensure that summaries reflect an equitable representation of the input groups. This process involves embedding documents using BERT, dividing the dataset into fairlets, and applying k-median clustering to construct a diversity-preserving summary.

The formal descriptions of the algorithm are presented in Algorithm 2.

Algorithm 2 FairExtract Algorithm

Input:

- Document set $\mathcal{D}$ of size N
- Groups G_1 and G_2
- Proportions g_1 (for G_1) and g_2 (for G_2) where $\gcd(g_1, g_2) = 1$
- Desired summary length L, where $L \ll N$

Output:

- Diversity-preserving extractive summary $\mathcal{S}$

Step 1: Embedding Documents
Embed each document $d_i \in \mathcal{D}$ into a vector in $\mathbb{R}^{768}$ using BERT.

Step 2: Fairlet Decomposition
Decompose $\mathcal{D}$ into fairlets, each containing g_1 documents from G_1 and g_2 from G_2, minimizing the sum of Euclidean distances.

Step 3: Finding Fairlet Centers
For each fairlet, select the document that minimizes the sum of distances to other documents.

Step 4: k-Median Clustering on Fairlet Centers
Calculate $k = \frac{L}{g_1+g_2}$ and perform k-median clustering on the fairlet centers.

Step 5: Summary Construction
From each cluster, select the fairlet corresponding to the cluster center and add all documents from that fairlet to the final summary $\mathcal{S}$.

Return: The final summary $\mathcal{S}$

A.2 Sample of Dataset

Table 4 presents a sample of the dataset used in this study, containing tweets from different social groups about Chicago. Each entry indicates the group (White, African-American (AA), or Hispanic(Hisp)) and the corresponding tweet.

A.3 Impact of Varying Fairness Weight on Composite Metrics

In this section, we present the results of an experiment where we varied the weight assigned to fairness in the composite metric formula. Specifically, we explored the performance of FairExtract and FairGPT under different fairness weights to assess their robustness in balancing quality and fairness. Table 5 summarizes the results for the setting where the fairness weight α is reduced to 0.16, representing a 16% incentive toward fairness and an 84% incentive toward quality.

Group	Tweet
White	Turns out not all White Castles are the same. Why do you push me away Chicago?!
AA	"I mean I'm from Chicago. I'll cheer for the Bears, but I'm a bigger 49ers fan."
White	Nothing makes me happier than seeing the Bulls win _____ #ChicagoBasketball #Bullieve
White	If you see on the news something about the Chicago Kitchen Clown Bandits, then it will be referring to me, my friend Eten, and I.
AA	Truuu we tryna find sum to do too.. I dnt wanna b n Chicago if ain't nobody here.
White	Oh yeah.. I'm good. Hangin' up here in Chicago today. :)
Hisp	You girls have a safe flight.! See you in Chicago (:
...	**... (Dataset continues with more examples)**

Table 4: Sample of tweets from different social groups in the dataset. The full dataset contains many more examples.

Clustering-based Methods					
Model	SUPERT+F	BLANC+F	SumQA+F	BARTSc+F	UniEval+F
`Naive`	0.485	0.525	0.288	0.699	0.343
`NaiveFair`	0.530	0.578	0.337	0.744	0.373
`TextRank Vanilla`	0.488	0.397	0.335	0.687	0.323
`TextRank Cluster-A`	0.488	0.390	0.313	0.686	0.283
`TextRank Cluster-H`	0.491	0.394	0.321	0.672	0.285
`BERT-EXT Vanilla`	0.515	0.529	0.298	**0.756**	0.338
`BERT-EXT Cluster-A`	**0.539**	0.538	0.309	0.744	0.355
`BERT-EXT Cluster-H`	0.536	0.511	0.299	0.746	0.315
`FairExtract (Ours)`	0.537	**0.593**	**0.339**	0.740	**0.396**
LLM-based Methods					
`ChatGPT-EXT`	**0.764**	0.545	0.288	0.899	0.396
`FairGPT (Ours)`	0.726	**0.597**	**0.354**	**0.907**	**0.446**

Table 5: Evaluation results using composite metrics for clustering-based and LLM-based summarization methods with reduced fairness weighting ($\alpha = 0.16$). The best values for each metric are highlighted in bold.