Fair Summarization:
Bridging Quality and Diversity in Extractive Summaries

Sina Bagheri Nezhad, Sayan Bandyapadhyay, Ameeta Agrawal
Department of Computer Science
Portland State University, USA

Introduction

- Fairness in multi-document summarization is a critical challenge.
- Fairness is defined as achieving balanced representation of social groups in the summary, ensuring that each group is proportionally reflected as in the original dataset.
- This research introduces two methods for fair extractive summaries:
  - **FairExtract**: Clustering-based approach using k-median clustering with fairlet decomposition.
  - **FairGPT**: Uses GPT-3.5-turbo with fairness constraints.
- Goal: Ensure equal representation across dialects.

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?

Comparison of summaries. Tweets from different groups are highlighted: White and African-American. `

Research Questions

- How does achieving perfectly fair summaries impact overall quality?
- How do current methods balance both fairness and summarization quality?

Proposed Methods

FairExtract:

- Uses fair clustering with fairlet decomposition.
- Ensures proportional representation across social groups.
- Steps include document embedding, fairlet decomposition, and k-median clustering.

Algorithm 1 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 $\mathcal{S}$

FairGPT:

- Leverages LLM for fairness.
- Ensures balanced sentence selection from different dialect groups.
- Uses longest common subsequence to match generated summaries with original content.

Algorithm 2 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 the Prompt 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 $\mathcal{S}$.

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}"
```

Dataset & Evaluation

Dataset: DivSumm dataset with tweets from three social groups (White-aligned, Hispanic, African-American).
Metrics: Evaluation is conducted on pairwise group combinations, assessing Quality, Fairness, and composite metrics (Quality + Fairness).
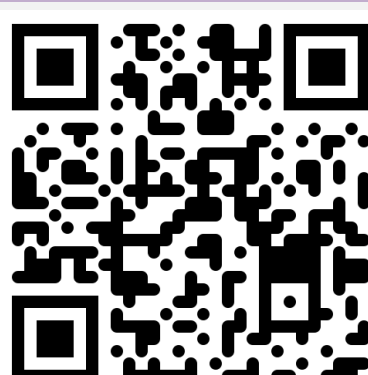**Baselines:** A variety of clustering and LLM-based methods.

Key Results

- FairExtract and FairGPT ensure perfect fairness while maintaining competitive summarization quality.
- Overall, FairGPT emerges as the best model, effectively balancing both quality and diversity across summaries.
- FairExtract performs is competitive across a wide range of metrics.
- Composite Metrics like SUPERT+F provide nuanced insights into fairness-quality trade-offs.

Model	SUPERT+F	BLANC+F	SumQA+F	BARTSc+F	UniEval+F
Clustering-based Methods					
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**

Takeaway

- Fairness in summarization is achievable without sacrificing quality.
- **FairExtract** and **FairGPT** offer promising methods for fair representation in summaries.
- Composite metrics provide a comprehensive view of the trade-offs between fairness and quality, helping to evaluate balanced summarization methods effectively.

Scan me

Acknowledgments