Understanding Position Bias Effects on Fairness in Social Multi-Document Summarization

Olubusayo Olabisi and **Ameeta Agrawal**
Portland State University
{oolabisi,ameeta}@pdx.edu

Abstract

Text summarization models have typically focused on optimizing aspects of quality such as fluency, relevance, and coherence, particularly in the context of news articles. However, summarization models are increasingly being used to summarize diverse sources of text, such as social media data, that encompass a wide demographic user base. It is thus crucial to assess not only the quality of the generated summaries, but also the extent to which they can fairly represent the opinions of diverse social groups. Position bias, a long-known issue in news summarization, has received limited attention in the context of social multi-document summarization. We deeply investigate this phenomenon by analyzing the effect of group ordering in input documents when summarizing tweets from three distinct linguistic communities: *African-American* English, *Hispanic-aligned* Language, and *White-aligned* Language. Our empirical analysis shows that although the textual quality of the summaries remains consistent regardless of the input document order, in terms of fairness, the results vary significantly depending on how the dialect groups are presented in the input data. Our results suggest that position bias manifests differently in social multi-document summarization, severely impacting the fairness of summarization models.

1 Introduction

As the use of natural language processing models gets more prevalent in various industries, academic and social settings, it is imperative that we assess not only the quality of these models but also their fairness when exposed to data originating from diverse social groups (Czarnowska et al., 2021). Text summarization models, in particular, facilitate the processing of large collections of a wide variety of text data by distilling documents into short, concise, and informative summaries while preserving the most relevant points from the source document (Nallapati et al., 2017; Zhang et al., 2018; Liu and Lapata, 2019). Multi-document summarization (MDS) is the task of generating a coherent summary from a set of input documents, usually centered around a topic, as opposed to single document summarization (SDS) which takes one document as input. The input in MDS consists of multiple documents, that may have been written by distinct users, varying in linguistic diversity, styles, or dialects.

MDS can be of type *extractive*, where the models extract the salient points directly from the source document to form the summary, or of type *abstractive* where the models generate summaries by rewriting salient information using novel words or phrases. In both cases, the resulting summary should be of good quality in terms of informativeness, coherence and relevance to the source document. At the same time, a good summary should be *unbiased* and should reflect the diversity of thoughts and perspectives present in the source documents.

The notion of fairness describes equal or fair treatment without favoritism or discrimination. However, plenty of evidence suggests intrinsic societal biases in language models (Bolukbasi et al., 2016; Bommasani et al., 2021; Deas et al., 2023). More specific to the task of summarization, fairness is measured by the ability of algorithms to capture the peculiarity in all represented groups (Shandilya et al., 2018; Dash et al., 2019; Keswani and Celis, 2021; Olabisi et al., 2022; Ladhak et al., 2023).

Conventionally, the documents in MDS are simply concatenated into one large collection of text as the input for the model. Prior research supports the existence of position bias, or lead bias, where the models rely excessively on the position of the sentences in the input rather than their semantic information (Lin and Hovy, 1997; Hong and Nenkova, 2014; Wang et al., 2019). This is a particularly common phenomenon in news summarization, where early parts of an article often contain the most salient information. While many

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 117–129

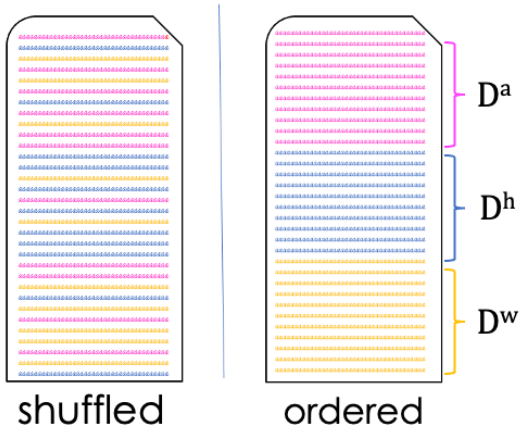


Figure 1: Illustration showing `shuffled` vs. `ordered` input for multi-document summarization consisting of documents from three diverse groups ($\mathcal{D}^a$, $\mathcal{D}^h$, $\mathcal{D}^w$) as indicated by the three colors. The `ordered` input is denoted as $\mathcal{O}^a$ when $\mathcal{D}^a$ documents appear first in the input.

algorithms exploit this fact in summary generation, it can have a detrimental effect when important information is spread throughout the input.

In non-news domains, weak or no position bias has been observed (Kedzie et al., 2018; Kim et al., 2019). Regardless of whether position bias is noted or not, previous investigations have quantified the *effects* of position bias mostly in terms of standard summarization metrics (e.g., ROUGE) which focus on the textual quality of the summary (Sotudeh et al., 2022; Scirè et al., 2023). In this work, we investigate the effects of position bias on the fairness of the generated summaries.

Specifically, we ask two questions: *(i)* Do the system summaries show any position bias when we vary the order of the input documents? *(ii)* What is the impact of position bias on the fairness of the system summaries?

For our experiments we use DivSumm, a summarization dataset of linguistically diverse communities representing three dialect groups (Olabisi et al., 2022). We explore the effects of position bias in the outputs of seven abstractive summarization models (and three extractive models) and under two investigation setups: `shuffled` (when the data is presented as randomly shuffled) and `ordered` (when the input documents are grouped according to their dialects). Figure 1 presents a schematic overview. The generated summaries are evaluated in terms of fairness, as well as metrics related to the textual quality.

The contributions of our work are as follows:

- We comprehensively investigate the phenomenon of position bias in the context of social multi-document summarization;
- We explore ten different summarization models, both abstractive and extractive;
- We contextualize and quantify the impact of position bias in terms of fairness and textual quality of generated summaries.

2 Related Work

In this section we present some notable prior research in two relevant areas. First, we discuss position bias in summarization, followed by works studying fairness in summarization.

Position Bias in Summarization Position bias can manifest in MDS scenarios just as it does in SDS scenarios because in MDS, the documents are typically concatenated into one long input and treated very much like a 'single' document. Several works have studied the substantial position bias (also known as lead bias), especially in the context of news summarization where the datasets and models prioritize selecting sentences from the beginning of an article (Lin and Hovy, 1997; Hong and Nenkova, 2014; Wang et al., 2019). Often the lead bias is so strong that the simple lead-k baseline or using the first k sentences of a news article to generate the summary can score higher than many other models (See et al., 2017). While some have suggested approaches for mitigating or countering lead bias (Grenander et al., 2019; Xing et al., 2021; Gong et al., 2022; Zhang et al., 2022), others have leveraged lead bias (Yang et al., 2020; Zhu et al., 2020; Padmakumar and He, 2021).

Interestingly, although position bias dominates the learning signal for news summarization or similar domains, it is less apparent in other domains where most non-news datasets show weak or no position bias (Kedzie et al., 2018; Jung et al., 2019; Kim et al., 2019; Sharma et al., 2019; Sotudeh et al., 2022; Scirè et al., 2023). Notably, none of these studies consider datasets where data originates from diverse social groups, which is the focus of our work.

Moreover, prior research studying the effect of position bias has quantified its impact exclusively in terms of textual quality, typically measured in terms of summarization metrics such as ROUGE, and others. To our knowledge, ours is the first work quantifying the impact of position bias in

multidocument summarization in terms of fairness where data originates from diverse social groups.

Fairness in Summarization A significant amount of work has been done toward improving the textual quality of summaries but not so much in terms of enhancing the fairness of summaries, particularly in the context of diverse groups. Prior text summarization work has proposed fairness-preserving algorithms (Shandilya et al., 2018; Dash et al., 2019), bias mitigation models (Keswani and Celis, 2021) and fairness interventions for extractive and abstractive summarization (Olabisi et al., 2022). Furthermore, Ladhak et al. (2023) observed that name-nationality stereotypes propagate from pretraining data to downstream summarization systems and manifest as hallucinated facts.

3 Experimental Setup

Considering the extensive literature on fairness in natural language processing, which highlights significant disparities in the processing of data from different social groups, whether along the dimensions of gender or race or others, we are compelled to ask two questions:

1. What happens when the input data to be summarized is deliberately grouped according to the social groups, such as dialect groups in our case? (in Section 4) and,

2. How do the effects of position bias affect the fairness of generated summaries (Section 5).

Before exploring these questions, we first describe our experimental setup in this section.

3.1 Task Formulation

Considering a multi-document set of n topically-related documents $\mathcal{D} = \{d_1^{g_1}, ..., d_n^{g_r}\}$, where each document belongs to one of several diverse social groups $\mathcal{G} = \{g_1, ..., g_r\}$, the objective is to produce a summary $\mathcal{S}(\mathcal{D})$ that ideally exhibits both high textual quality and fairness. In this work, because of the original dataset design where the number of documents from each group is equal in the input, our investigation is concerned with the notion of equal representation. As such, a summary is considered to be fair when all groups $g_1, ..., g_r$ are equally represented in the output.

3.2 Dataset

For our experiments, we use the DivSumm dataset[1], an MDS dataset consisting of English tweets of three diverse dialects (*African-American* English, *Hispanic-aligned* Language, and *White-aligned* Language) (Olabisi et al., 2022), which was developed using a large corpus of tweets originally collected by Blodgett et al. (2016). The dataset includes 25 topically-related sets of documents (tweets) as input and corresponding human-written extractive and abstractive summaries. Each set $\mathcal{D}$ consists of 90 documents evenly distributed among the three dialects (i.e., 30 documents per dialect). A selection of dialect diverse tweets from DivSumm is presented in Table 3.

3.3 `Shuffled` and `Ordered`

To study the phenomenon of position bias in social multi-document summarization where documents originate from different social groups, we devise two distinct scenarios: `shuffled` and `ordered`, as depicted in Figure 1.

In the **shuffled** setting, documents appear randomly present in the input in no specific order. In fact, to ensure consistency, we retain the original order as presented in the DivSumm dataset which the annotators used to craft the summaries.

In the **ordered** setting, we perturb the input data by grouping documents from each social group together. When the subset of *White-aligned* Language tweets ($\mathcal{D}^w$) appears first, the input set is denoted as $\texttt{ordered}^{\texttt{white}}$ or, simply, $\mathcal{O}^w$. Similarly, when the subset of *African-American* English tweets ($\mathcal{D}^a$) come first, we denote that set as $\mathcal{O}^a$, and when the subset of *Hispanic-aligned* Language documents ($\mathcal{D}^h$) appears first, we denote that set as $\mathcal{O}^h$. Specifically, the input documents are ordered as follows:

$$\mathcal{O}^w = \{\mathcal{D}^w, \mathcal{D}^a, \mathcal{D}^h\}$$
$$\mathcal{O}^a = \{\mathcal{D}^a, \mathcal{D}^h, \mathcal{D}^w\}$$
$$\mathcal{O}^h = \{\mathcal{D}^h, \mathcal{D}^w, \mathcal{D}^a\}$$

These documents are summarized using several models described in the next section, allowing us to subsequently investigate the different summaries we generate – $\mathcal{S}(\mathcal{O}^w)$, $\mathcal{S}(\mathcal{O}^a)$, $\mathcal{S}(\mathcal{O}^h)$, and $\mathcal{S}(\texttt{shuffled})$ – which are obtained from four distinct sets of input documents – $\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$, and `shuffled`, respectively.

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

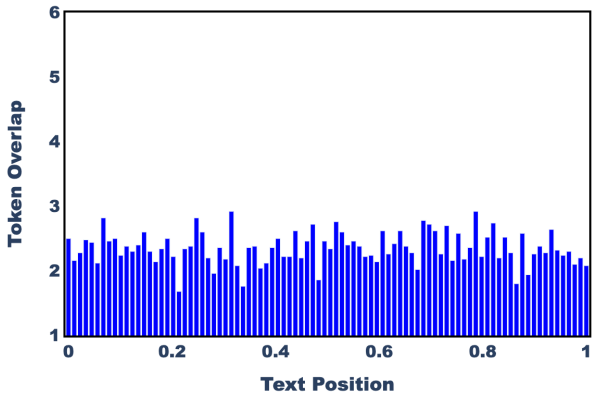


Figure 2: Average token overlap between human-written reference summaries and each document d_i using the DivSumm dataset. Text position on the x-axis has been normalized between 0 and 1.

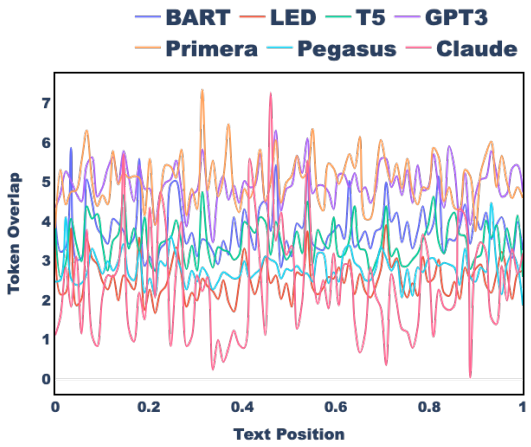


Figure 3: Average token overlap between `ordered` system-generated summaries by each abstractive summarization model and each document d_i in the input set $\mathcal{D}$ of DivSumm. Text position on the x-axis has been normalized between 0 and 1.

3.4 Summarization Models

We study a total of seven abstractive models in our experiments. We also study three extractive models, the details and results of which are discussed in A. Following the setup of DivSumm, we generate summaries of 5 sentences per topic

The seven abstractive models included in our experiments are as follows:

- BART[2] (Lewis et al., 2019),
- T5 (Raffel et al., 2019),
- LED (Longformer Encoder-Decoder) (Beltagy et al., 2020),
- PEGASUS (Zhang et al., 2020),
- GPT-3.5,
- PRIMERA (Xiao et al., 2021), and
- CLAUDE (`Claude 3 Opus`).

GPT-3.5 and Claude were prompted with the following prompt – "Please summarize the following texts in only five sentences".

4 Position Bias in Social MDS

This section discusses position bias within three types of summaries: human-authored reference summaries of the DivSumm dataset, system summaries generated using the `shuffled` input, and system summaries generated using `ordered` inputs.

[2]Model checkpoints for BART, T5, LED, Pegasus, and Primera were accessed from `https://huggingface.co/models`.

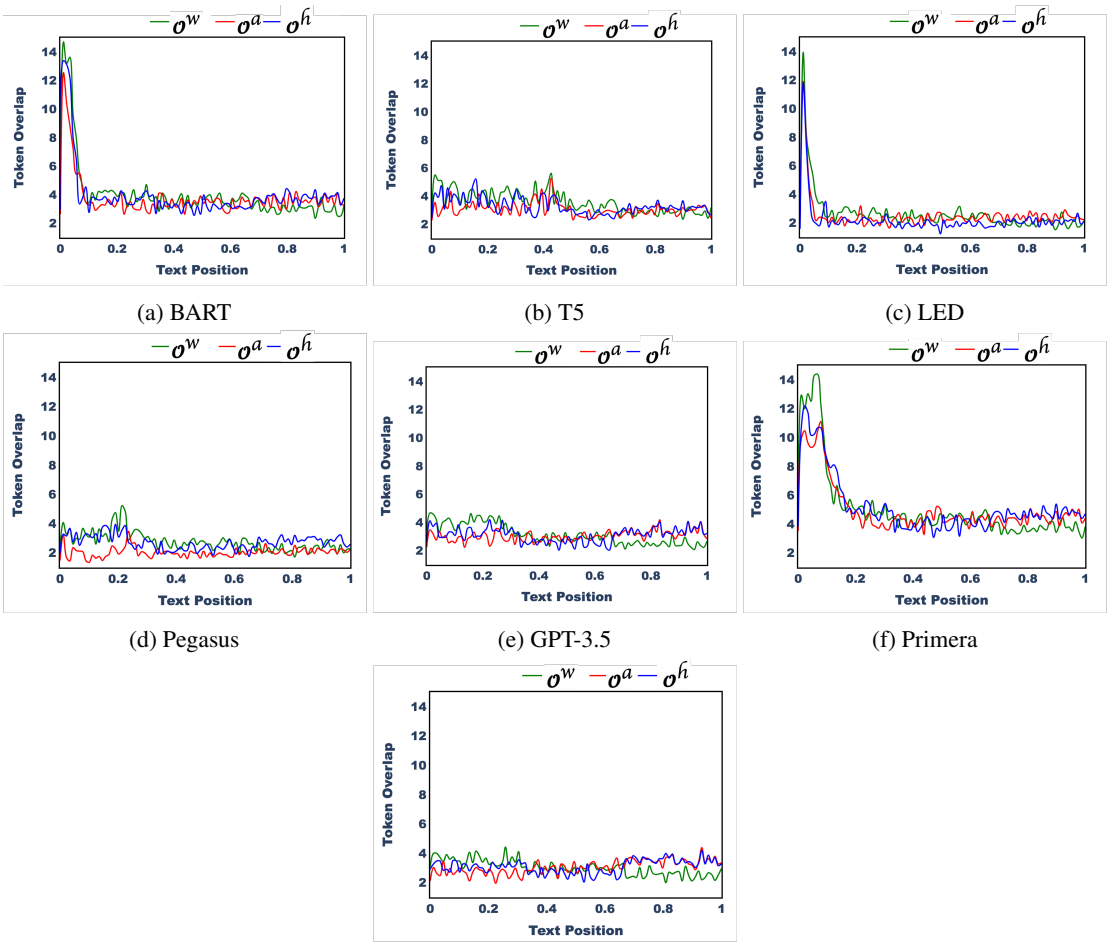
Following prior work on position bias, we calculate the overlap between the summaries and the input documents by computing the number of tokens shared between the summary and each document of the MDS topic set. That is, given the 90 documents in each topically-related input set, we get the overlap score for each document $(d_1, d_2, ..., d_{90})$ with respect to a summary, and report the average score over the entire dataset. A higher overlap score implies more semantic relationship between the summary and source document.

4.1 Position Bias in Human-Written Reference Summaries

To examine position bias in the summaries created by humans, we analyze both abstractive and extractive reference summaries of DivSumm dataset. Because the dataset contains two reference summaries per input, we report the average score. The results are presented in Figure 2 where no noticeable position bias is observed, and it is encouraging to note that the annotators were not influenced by the position of the documents in the input when producing their summaries.

4.2 Position Bias in System Summaries (`Shuffled`)

The results of position bias within model-generated summaries using `shuffled` inputs are presented in Figure 3. Similar to the human-written reference summaries, we observe no notable position bias suggesting that when summarizing randomly shuf-



(a) BART (b) T5 (c) LED

(d) Pegasus (e) GPT-3.5 (f) Primera

(g) Claude

Figure 4: Average token overlap between `ordered` system-generated summaries by each of the seven abstractive summarization models and each document d_i in the input set $\mathcal{D}$ of the DivSumm dataset. Text position on the x-axis has been normalized between 0 and 1.

fled data from various social groups, the models also do not exhibit any particular lead bias. This observation on DivSumm, a dataset of tweets, is consistent with trends observed in other social datasets (Reddit posts (Kim et al., 2019) and social user posts (Sotudeh et al., 2022)).

4.3 Position Bias in System Summaries (`Ordered`)

Now we discuss the results of position bias in system summaries that were generated using various `ordered` inputs: $\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$. Model-specific results are presented in Figure 4, where, interestingly, we now observe a **strong position bias in three out of seven abstractive models,** (BART, LED, and Primera), with up to 3 times higher token overlap in the beginning of the input document, as shown by the distribution. Three other models show weak position bias (T5, Pegasus, and GPT-3.5). This phenomenon diverges from traditional position bias, where models tend to favor earlier bits of text. *Instead, we notice that models favor earlier pieces of text only when the text exhibits some socially linguistic similarity*. These observations highlight the importance of more nuanced analysis when exploring position bias in summarization systems, especially when processing diverse social data.

5 Fairness and Textual Quality Amidst Position Bias

Having observed an instance of position bias, especially when input data is grouped according to dialect groups, the next natural question to ask is how does this position bias quantitatively impact the fairness and textual quality of the generated summaries. We briefly describe the evaluation metrics before discussing the main results.

Model	$\mathcal{O}^w$				$\mathcal{O}^a$				$\mathcal{O}^h$				shuffled			
	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)
BART	**0.64**	0.41	0.45	0.23	0.41	**0.55**	0.40	0.15	0.44	0.42	**0.59**	0.17	**0.41**	**0.41**	0.40	0.01
LED	**0.47**	0.30	0.33	0.17	0.31	**0.43**	0.31	0.12	0.26	0.24	**0.36**	0.12	0.30	0.29	**0.35**	0.06
T5	**0.52**	0.39	0.48	0.13	0.39	**0.46**	0.43	0.07	0.40	0.41	**0.49**	0.09	0.37	**0.41**	0.40	0.04
PEGASUS	**0.34**	0.28	0.29	0.06	0.22	**0.25**	0.21	0.04	0.26	0.24	**0.32**	0.08	**0.32**	0.33	**0.32**	0.01
GPT-3.5	**0.47**	0.35	0.38	0.12	**0.38**	**0.38**	0.36	0.02	0.38	0.34	**0.41**	0.07	**0.40**	0.35	0.37	0.05
PRIMERA	**0.62**	0.41	0.45	0.21	0.42	**0.60**	0.44	0.18	0.45	0.44	**0.62**	0.18	**0.49**	0.48	0.50	0.02
CLAUDE	**0.39**	0.33	0.36	0.06	**0.37**	0.32	0.34	0.05	**0.36**	0.31	0.34	0.05	**0.37**	0.32	0.35	0.05
AVG	**0.49**	0.35	0.39	0.14	0.36	**0.43**	0.36	0.09	0.36	0.35	**0.45**	0.11	0.38	0.37	0.39	0.04

Table 1: ***Fairness.*** Similarity scores of summaries generated by `ordered` inputs ($\mathcal{O}^w, \mathcal{O}^a, \mathcal{O}^h$) and `shuffled` inputs compared to each group of documents ($\mathcal{D}^w$, $\mathcal{D}^a$, $\mathcal{D}^h$) across seven abstractive summarization models using the *DivSumm* dataset. The highest similarity scores are shown in bold.

5.1 Evaluation Metrics

Fairness (Gap): One way of measuring fairness is by estimating the amount of representation from each dialect group in the final summary by comparing the summary $\mathcal{S}$ to the set of documents from each group. Given that an unbiased summary should capture the perspectives across all groups, we evaluate summary fairness for both extractive and abstractive models using semantic similarity of the summary to each represented group. As an example, for input $\mathcal{O}^w$, we compare the final summary $\mathcal{S}(\mathcal{O}^w)$ to the document set of each dialect group: $\mathcal{D}^w$, $\mathcal{D}^w$, and $\mathcal{D}^h$. In other words, we compute $sim(i, j)$ where $i = \{\mathcal{S}(\mathcal{O}^w), \mathcal{S}(\mathcal{O}^a), \mathcal{S}(\mathcal{O}^h)\}$ and $j = \{\mathcal{D}^w, \mathcal{D}^a, \mathcal{D}^h\}$. Similarity can be estimated by many possible methods of obtaining semantic similarity. We use cosine similarity.

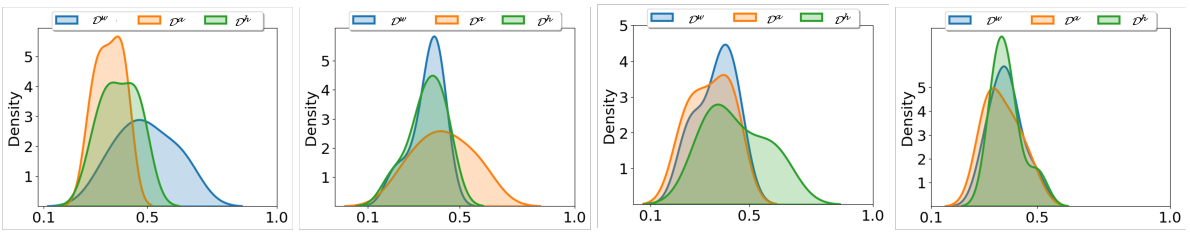
From these similarity scores, we can derive the **Fairness Gap (ΔFair)** by calculating the difference between the maximum and the minimum scores attributed to any of the groups (Olabisi et al., 2022). Intuitively, a summary that produces relatively similar representation scores across all groups can be considered as *fair* because it likely contains comparable representation from all groups such that no one group is significantly underrepresented.

Textual Quality: Four established metrics are used for assessing the quality of the summaries: ROUGE, BARTScore, BERTScore, and UniEval. **ROUGE** (Lin, 2004) calculates the lexical overlap between the model-generated summary and the reference summaries. For our experiments, we report the F1 scores of ROUGE-L which is the longest common subsequence between the two summaries. **BARTScore** (Yuan et al., 2021) leverages BART's average log-likelihood of generating the evaluated summary conditional on the source document. Since it uses the average log-likelihood for target tokens, the calculated scores are smaller than 0 (negative). We use the `facebook/bart-large-cnn` checkpoint. **BERTScore** (Zhang* et al., 2020) relies on BERT embeddings and matches words in system-generated summaries and reference summaries to compute token similarity. We use the `microsoft/deberta-xlarge-mnli` model and report the F1 scores. **UniEval** (Zhong et al., 2022) is a unified multi-dimensional evaluator that employs boolean question answering format to evaluate text generation tasks. We make use of `unieval-sum` which evaluates system-generated summaries in terms of four dimensions: coherence, consistency, relevance and fluency. Except for fluency, the rest are reference-free metrics. We report the overall score.

5.2 Results

Evaluating fairness. The results in Table 1 report the fairness scores for all seven models. **We clearly observe that ordering the input documents based on groups certainly favors the group that appears first.** This phenomenon is consistently observed in all three types of `ordered` sets, regardless of which particular dialect group's data is presented first. However, when the documents are presented as `shuffled`, no single group is over-represented and the summaries appear more balanced (ΔFair = 0.04).

The density plots in Figure 5 also show that the `shuffled` input set is the most balanced across all groups, unlike the `ordered` sets which are significantly skewed. Furthermore, amongst `ordered`



(a) $\mathcal{O}^w$ (b) $\mathcal{O}^a$ (c) $\mathcal{O}^h$ (d) shuffled

Figure 5: Density distribution of similarity scores between system-generated summaries and each group, across all summarization models for $\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$ and shuffled input sets. The outputs of shuffled inputs show very different and balanced distributions compared to the ordered inputs.

Model	ROUGE-L				BARTSCORE				BERTSCORE				UNIEVAL			
	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.
BART	**0.15**	0.14	0.14	**0.15**	-3.73	-3.74	-3.72	**-3.69**	**0.51**	0.50	**0.51**	0.50	0.46	0.46	**0.48**	0.44
T5	**0.15**	0.13	0.13	0.14	-3.76	-3.75	-3.74	**-3.72**	0.50	0.48	0.49	**0.51**	0.45	0.45	**0.47**	0.44
LED	**0.12**	0.11	0.10	**0.12**	-3.75	-3.79	-3.79	**-3.73**	0.44	0.40	0.39	**0.47**	0.44	0.44	**0.46**	0.43
PEGASUS	**0.14**	0.11	0.13	**0.14**	**-3.73**	-3.75	-3.76	**-3.73**	**0.47**	0.44	0.46	0.46	0.45	0.45	**0.47**	0.43
GPT-3.5	0.20	0.20	**0.21**	**0.21**	-3.64	-3.68	**-3.62**	-3.65	0.57	0.58	**0.59**	**0.59**	0.46	0.45	**0.48**	0.44
PRIMERA	**0.14**	0.12	0.13	0.13	-3.67	-3.68	**-3.63**	-3.64	0.51	0.49	**0.50**	0.49	0.45	0.46	**0.48**	0.44
CLAUDE	0.18	0.18	**0.19**	0.18	**-3.64**	**-3.64**	**-3.64**	-3.65	0.56	0.56	**0.57**	0.56	0.44	0.44	**0.46**	0.43
AVG	0.15	0.14	0.15	0.15	-3.70	-3.72	-3.70	-3.69	0.51	0.49	0.50	0.51	0.45	0.45	0.47	0.44

Table 2: ***Quality***. Results of ordered ($\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$) and shuffled (Sh.) approaches across seven abstractive summarization models showing ROUGE-L, BARTScore, BERTScore, and UniEval scores on the *DivSumm* dataset. The best scores are shown in **bold**, whereas the highest scores per metric are shown as underlined.

documents, the fairness gap is the largest when documents of White-aligned language are passed first (ΔFair = 0.14), and the smallest when documents of African-American English appear first (ΔFair = 0.09).

Evaluating textual quality. Table 2 presents the summary quality scores across all seven summarization models for the four sets of input. We clearly see that the scores of the shuffled approach are superior or comparable to the scores from the three input sets in the ordered approach, except in the case of UniEval. **This shows that with respect to quality, there is no significant difference whether documents are presented as ordered or shuffled.**

5.3 Discussion

Some samples of system summaries are presented in Table 3. The key findings of our study can be summarized as follows:

- We find no evidence of position bias in human-annotated reference summaries of DivSumm, a social MDS dataset of diverse groups. Same observation is made for the abstractive system-generated summaries obtained when the input documents are passed in randomly or shuffled.
- However, when the input is ordered based on dialect groups, we observe a significant position bias in the system summaries, with the summaries having higher overlap with the group that appears first in the input document.
- Ordered documents involving different dialects result in summaries that are significantly skewed in terms of fairness, with the group whose data appears first is clearly favored by the models. In contrast, shuffled documents show the least amount of fairness gap.
- In terms of quality, we observe that for all models and metrics, the scores for ordered and shuffled remain comparable, suggesting that ordering based on diverse groups has no noticeable effect on the quality of system-generated summaries.

Taken together, the findings of our study indicate that both the ordered and shuffled approaches yield comparable results in terms of textual quality, but highly disparate results in terms of fairness. This phenomenon is consistently observed in

Input Documents Set	
d_1: **Hispanic** : The Grammys should have come out on Saturday so I won't stay up late today lol d_2: **AA** : Wasn't it during the Grammys the last time Chris Brown slid Rhianna? d_3: **White** : Feel free to join my lonesome self swimminngg at Grammys!! d_4: **AA** : I've given up #DowntonAbbey for J.T.? This is serious #Grammys d_5: **Hispanic** : oh lol thanks thought you were talking about the Grammys lol sorry lol d_6: **Hispanic** : I don't even know if I am watching the right latin Grammys lol d_7: **White** : "If I'm a hipster about anything, it's Kings of Leon. I listened to them before they won Grammys." d_8: **White** : isn't performing at the Grammys? What's the point of even having the Grammys now? d_9: **AA** : Imma get some ideas for Easter from these here Grammys . ____ d_{10}: **Hispanic** : Wow every celeb in the crowd is singing Bruno's lyrics. Awesome. #Grammys ... ... d_{88}: **AA** : These Grammys need some life in it cause I'm so bored d_{89}: **AA** : Got my easter outfit nd dress for the Grammys.. im set :) d_{90}: **White** : "great, now she can have practice for whe she doesn't win at next year's Grammys"	
Model	**Summary**
BART	I've given up DowntonAbbey for J.T.? This is serious Grammys. The Grammys lasted an extra 30 mins just for us to all get hazed by LL smh. Im not even watching the Grammys im too lazy to change the channel but isn't Justin doing a livestream or something at the same time. I'm gonna say yes.
GPT-3.5	Tina Campbell won't be attending the Grammys, and J Cole is being tipped for success. The Black Keys won five awards at the Grammys, and Eminem is expected to win Best Male Singer and Rapper. Rihanna and Chris Brown were together at the Grammys, and there is a lot of Twitter beef happening due to the event. An ad agency has the Grammys as their client, and a speech teacher asked someone to dress for the Grammys for a show. People are discussing the Grammys and celebrities attending.

Table 3: Example of a shuffled input set from the *DivSumm* dataset showing input documents with their respective dialects, and system summaries generated by the model with the lowest fairness gap (BART) and highest textual quality (GPT-3.5). Note that the dialect labels shown here are only for clearer presentation and are not part of the input to the model.

all abstractive models, suggesting that the models are not robust to fairly straightforward group-level data perturbations. These findings are important because they highlight a potential source of nuanced bias in the summarization models. The observation that ordering the input documents based on groups favors the group that appears first indicates a systematic bias in the models' behavior. The fact that the shuffled input set leads to more balanced summaries across all groups implies that the bias observed in the ordered sets can be mitigated by introducing randomness in the presentation of input data. This insight is crucial for understanding and addressing bias in summarization systems, especially in scenarios where fairness and equity are important considerations, such as in social data analysis or decision-making processes. Overall, this result sheds light on an important aspect of model behavior and informs strategies for improving the fairness and effectiveness of summarization models.

6 Conclusion

In this work, we investigate how position bias manifests in social multi-document summarization, specifically in scenarios where the input data is derived from three linguistically diverse communities. When presented with randomly shuffled input data, summaries generated by ten distinct summarization models exhibited no signs of position bias. However, a significant shift occurred when the input data was simply reordered based on social groups. In such instances, the models produced biased summaries, primarily favoring the social group that appeared earlier in the input sequence. In terms of the quality of generated summaries, however, there was no notable difference due to the order in which source documents were presented, whether shuffled or ordered. Our results suggest that position bias manifests differently in the context of social multi-document summarization. Furthermore, they highlight the need to incorporate randomized shuffling in multi-document summarization datasets particularly when summarizing documents from diverse groups to ensure that the resultant summaries are not only of high quality but also faithfully representative of the diversity present in the input data.

Ethical Considerations

Our findings and conclusions in this paper are based on an existing social media summarization dataset composed of tweets in English, primarily

due to the lack of appropriate resources available to undertake such studies. Given the nature of naturally occurring data, it is possible that the data contains some offensive language. Hence, it is possible for the models to also generate summaries with offensive words. In addition to this, due to the constraint on tweet length, users are known to use acronyms and slangs that may have various meanings across different groups – this phenomenon is not accounted for in this study. Also, the existing dataset that we use in this work was originally collected from a corpus using geolocation and census data. This dialectal information used in categorizing users' languages should not be used as a representation of users' racial information. In this work, we evaluate summary fairness using proxy metrics such as semantic similarity to each represented group. The definition of fairness may vary for humans, and as such this should not be used as the gold standard.

Acknowledgments

We thank the anonymous reviewers as well as the members of PortNLP lab for their insightful comments. This research was supported by National Science Foundation grants (CRII:RI 2246174 and SAI-P 2228783).

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *arXiv preprint arXiv:2106.14574*.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 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.

Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shuai Gong, Zhenfang Zhu, Jiangtao Qi, Chunling Tong, Qiang Lu, and Wenqing Wu. 2022. Improving extractive document summarization with sentence centrality. *PloS one*, 17(7):e0268278.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. *arXiv preprint arXiv:1909.04028*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. *arXiv preprint arXiv:1908.11723*.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Vijay Keswani and L Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, pages 3802–3814.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *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 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Fifth conference on applied natural language processing*, pages 283–290.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2023. Echoes from alexandria: A large resource for multilingual book summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 853–867, Toronto, Canada. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. MentSum: A resource for exploring summarization of mental health online posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2682–2692, Marseille, France. European Language Resources Association.

Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 948–954, Online. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. 2022. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 127–141, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. *arXiv preprint arXiv:1808.07187*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2020. Make lead bias in your favor: Zero-shot abstractive news summarization. In *International Conference on Learning Representations*.

A Fairness in Extractive Models

We repeat the same experiments and analysis for extractive models to observe if they exhibit behavior similar to that observed in the abstractive models.

A.1 Summarization Systems

We study three summarization models in our experiments to generate summaries of 5 sentences per topic (multi-document set):

TEXTRANK[3] (Mihalcea and Tarau, 2004), an unsupervised graph-based ranking method, determines the most important sentences in a document based on information extracted from the document itself.

BERT-EXT[4] (Miller, 2019), an extractive summarization model built on top of BERT (Devlin et al., 2018), uses k-means clustering to select sentences closest to the centroid as the summaries.

LONGFORMER[5] (Beltagy et al., 2020) is a modification of the transformer architecture, using a self-attention operation that scales linearly with the sequence length.

A.2 Evaluation Metrics

In evaluating textual quality, We use the same four metrics used for the abstractive models. To estimate fairness (gap), in addition to semantic similarity used in evaluating the fairness of abstractive models, we consider **coverage** as well which measures the extent to which a summary is a derivative of the input text. Following previous literature (Dash et al., 2019; Keswani and Celis, 2021), we estimate group fairness via disparity in *extractive fragment coverage* (Grusky et al., 2018), which indicates the degree of surface-level text overlap by computing the percentage of words in the summary from each dialect group's collection of documents.

A.3 Results

While `shuffled` extractive models show no noticeable position bias in Figure 6, we observe a strong position bias using `ordered` inputs in two out of three extractive models (BERT and LongFormer), as shown in Figure 7 further highlighting the importance of exploring position bias in summarization of diverse social data.

[3] https://radimrehurek.com/gensim_3.8.3/summarization/summariser.html

[4] https://pypi.org/project/bert-extractive-summarizer/

[5] Model checkpoint for Longformer was accessed from https://huggingface.co/models

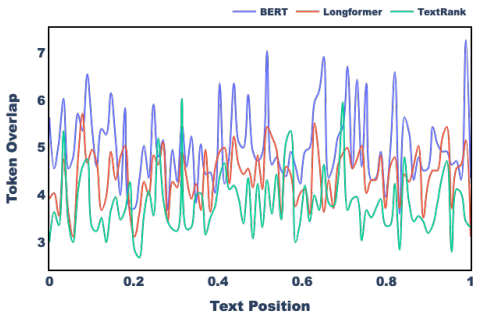
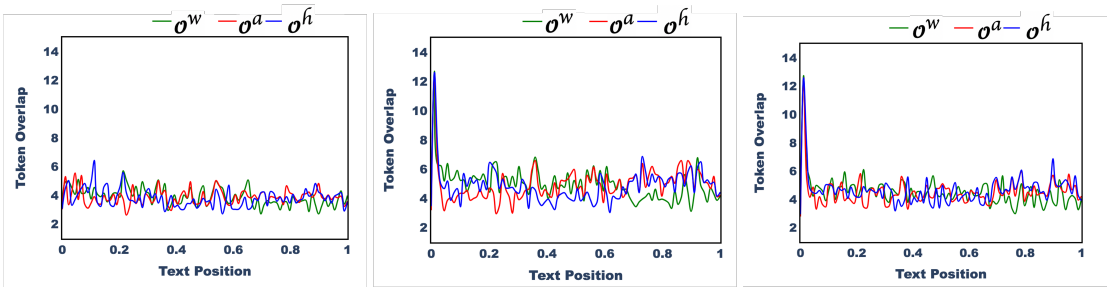


Figure 6: Average token overlap between `shuffled` system-generated summaries by each of the three extractive summarization models and each document d_i in the input set $\mathcal{D}$ of DivSumm. Text position on the x-axis has been normalized between 0 and 1.

Tables 4 and 5 show the fairness scores in terms of coverage and similarity, respectively, of extractive summaries. For all three models, we observe that the summaries generated using the `ordered` sets distinctly favor the group that appeared first in the input set of documents, while this phenomenon is absent from the `shuffled` set, where the results are much more evenly distributed across the three groups for all three models. Table 6 presents the quality scores along four metrics where, similar to abstractive models, little difference is noted between `ordered` and `shuffled` approaches.



(a) *TextRank* (b) *BERT* (c) *LongFormer*

Figure 7: Average token overlap between `ordered` system-generated summaries by each of the extractive summarization models and each document d_i in the input set $\mathcal{D}$ of DivSumm. Text position on the x-axis has been normalized between 0 and 1.

Model	$\mathcal{O}^w$				$\mathcal{O}^a$				$\mathcal{O}^h$				`shuffled`			
	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair
TEXTRANK	**0.80**	0.72	0.76	0.08	0.70	**0.81**	0.74	0.11	0.72	0.73	**0.82**	0.10	0.74	0.76	**0.78**	0.04
BERT	**0.78**	0.69	0.77	0.09	**0.75**	0.74	0.73	0.02	0.78	0.69	**0.80**	0.11	**0.77**	0.74	0.76	0.03
LONGFORMER	**0.77**	0.72	0.73	0.05	0.70	**0.80**	0.71	0.10	0.73	0.72	**0.79**	0.07	0.72	**0.78**	0.77	0.06
AVG	0.78	0.71	0.75	0.07	0.72	0.78	0.73	0.07	0.74	0.71	0.80	0.09	0.74	0.76	0.77	0.05

Table 4: ***Fairness.*** Coverage scores of `ordered` and `shuffled` approaches compared to each group of documents ($\mathcal{D}^w$, $\mathcal{D}^a$, $\mathcal{D}^h$) for three extractive summarization models on *DivSumm* dataset. The highest scores are shown in bold.

Model	$\mathcal{O}^w$				$\mathcal{O}^a$				$\mathcal{O}^h$				`shuffled`			
	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair
TEXTRANK	**0.57**	0.55	0.52	0.05	0.51	**0.54**	0.49	0.05	**0.55**	0.54	0.50	0.04	0.45	**0.46**	0.42	0.05
BERT	**0.61**	0.54	0.53	0.07	0.51	0.59	**0.61**	0.10	0.62	**0.63**	0.55	0.08	0.48	0.50	**0.52**	0.03
LONGFORMER	**0.58**	0.54	0.50	0.08	0.55	**0.56**	0.55	0.02	**0.54**	**0.54**	0.52	0.02	0.45	0.44	**0.47**	0.03
AVG	0.59	0.54	0.52	0.07	0.52	0.56	0.55	0.04	0.57	0.57	0.52	0.05	0.46	0.47	0.47	0.03

Table 5: ***Fairness.*** Semantic similarity scores of `ordered` and `shuffled` approaches compared to each group of documents ($\mathcal{D}^w$, $\mathcal{D}^a$, $\mathcal{D}^h$) across extractive summarization models on *DivSumm* dataset. The highest scores are shown in bold.

Model	**ROUGE-L**				**BARTSCORE**				**BERTSCORE**				**UNIEVAL**			
	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.
TEXTRANK	**0.23**	0.21	0.22	**0.23**	-4.42	-4.42	-4.44	**-4.29**	0.55	0.54	0.55	**0.56**	0.46	0.46	**0.48**	0.44
BERT	**0.24**	**0.24**	0.23	0.21	**-4.28**	-4.33	-4.39	-4.71	**0.56**	**0.56**	**0.56**	0.55	0.47	0.46	**0.49**	0.45
LONGFORMER	**0.22**	0.21	**0.22**	0.20	-4.38	-4.44	-4.41	**-4.35**	**0.56**	0.55	**0.56**	**0.56**	0.46	0.46	**0.48**	0.45
AVG	0.23	0.22	0.22	0.22	-4.36	-4.40	-4.41	-4.45	0.56	0.55	0.55	0.56	0.46	0.46	0.48	0.45

Table 6: ***Quality***. Results of `ordered` and `shuffled` approaches across extractive summarization models showing ROUGE-L, BARTScore, BERTScore and UniEval scores on *DivSumm* dataset. The best scores are shown in **bold**, whereas the highest scores per metric are shown as underlined.