Evaluating Summaries of Social Media Posts: Balancing Fairness for Readers and Authors

GARIMA CHHIKARA, Indian Institute of Technology Delhi, New Delhi, India and Delhi Technological University, Delhi, India
KRIPABANDHU GHOSH, Indian Institute of Science Education and Research Kolkata, Mohanpur, India
SAPTARSHI GHOSH, Indian Institute of Technology Kharagpur, Kharagpur, India
ABHIJNAN CHAKRABORTY, Indian Institute of Technology Kharagpur, Kharagpur, India

Text summarization has a wide range of applications, from condensing lengthy documents to summarizing user-generated content (such as tweets, Facebook posts, and Reddit discussions) that reflect diverse viewpoints on socially significant topics. Traditionally, summarization has been considered a reader-centric task, and algorithms are evaluated against human-written reference summaries using metrics such as the ROUGE score. However, this approach has two primary limitations when applied to user-generated content, especially written on sensitive issues. First, through surveys with participants from various countries, we demonstrate that human-written reference summaries can contain implicit biases. These biases introduce significant variations in the evaluation of summarization algorithms, raising concerns about their reliability in identifying summaries that effectively meet reader expectations. Second, we show that algorithmically generated summaries, as ranked by traditional evaluation methods, often fail to align with authors' preferences. These issues render the existing summary evaluation framework suboptimal for both readers and the authors of the content being summarized.

In this work, by reimagining summarization of user-generated content as a two-sided problem, we emphasize the importance of satisfying both readers and authors. Specifically, we take a two-pronged approach to tackle the issues with existing evaluation setup: (i) we introduce the concept of inequality in reader satisfaction to tackle the implicit biases in reference summaries; and (ii) we propose a novel evaluation framework based on satisfying the authors of the content. More specifically, we present an author satisfaction-based evaluation metric `CROSSEM` which, we show empirically, can complement the current summary evaluation paradigm. This work represents the first attempt to develop a fair summary evaluation framework for social media posts and is likely to spark future research in this domain.

CCS Concepts: • **Information systems** → **Summarization**; **Evaluation of retrieval results**.

Additional Key Words and Phrases: Fair Summarization, Summary Evaluation, Author Satisfaction

1 Introduction

In recent years, the amount of user generated information available on internet has increased tremendously. Specifically, social media has encouraged the participation of different communities in online discussions, resulting

This work is an extended version of the short paper by Chhikara et al [17], titled "Fairness for both Readers and Authors: Evaluating Summaries of User Generated Content", which was presented at the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) in 2023. The extension is detailed in Section 2.6.

Authors' Contact Information: Garima Chhikara, Indian Institute of Technology Delhi, New Delhi, India and Delhi Technological University, Delhi, India; e-mail: garimachhikara128@gmail.com; Kripabandhu Ghosh, Indian Institute of Science Education and Research Kolkata, Mohanpur, West Bengal, India; e-mail: kripa.ghosh@gmail.com; Saptarshi Ghosh, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India; e-mail: saptarshi@cse.iitkgp.ac.in; Abhijnan Chakraborty, Indian Institute of Technology Kharagpur, Kharagpur, West Bengal, India; e-mail: chakraborty.abhijnan@gmail.com.

ACM 1559-114X/2026/7-ART
https://doi.org/10.1145/3833377

in a data explosion. For example, X (formerly Twitter) receives around 500 million posts per day, in more than 50 different languages [77]. With such a large amount of data circulating in the digital space, there is a need to develop algorithms that can automatically shorten longer texts and deliver accurate summaries that fluently pass the intended messages. Such algorithms are classified as either *Extractive* or *Abstractive* summarizers. In extractive summarization, the task is to extract key sentences or phrases directly from the source text without altering the original wording. In contrast, abstractive summarization generates new sentences to convey the core meaning of the source text, often through rephrasing and paraphrasing.

A substantial body of prior research has been dedicated to the development of techniques for summarizing long documents [2, 24]. Historically, the focus of summarization research has been on creating concise and coherent summaries for individual documents or small collections of them, such as news articles. However, in recent years, summarization has been increasingly applied on different types of *user-generated text* (e.g., tweets, Facebook or Reddit posts) [42, 45, 58], where the task is to summarize short independent textual posts written by many *authors*. There exists a fundamental difference between the two summarization tasks, i.e., summarizing a single long document and generating summaries from large numbers (possibly hundreds or thousands) of user-generated texts. In the latter case, the input text can contain a much wider variety of opinions expressed by thousands of different authors, as compared to the case of summarizing one or a small set of long documents. Especially, if the user generated posts are on a polarizing topic such as politics, gender-related issues, religion, etc., then the posts can contain many mutually conflicting opinions, and it is natural to ask *whether the different opinions are fairly reflected in the summary* [80, 83]. This is an important question because summarization provides visibility to the posts/opinions selected in the summary, and multiple downstream applications relying on summarization further increase the exposure. For example, Google News shows a collection of tweets as part of the full coverage on a topic [35]; even Google search on a given topic displays tweets alongside the standard search results [41]; the Library of Congress only stores a selection of tweets as part of its archive while emphasizing the importance of giving voice to the common people [66].

The prevalent evaluation setup for text summarization involves comparing the *algorithmic summary* (generated by a summarization algorithm) with *reference summaries* written by human experts (also known as *gold-standard summaries*), using standard evaluation measures like ROUGE [53]. Reader is the one who will read the summary to obtain the synopsis of textual information. Reference summaries are manually created by summarizers acting as proxy for what the eventual *readers* would want from a summary. Thus, *evaluation metrics like ROUGE which make use of reference summaries essentially attempt to capture reader satisfaction.* An algorithmic summary which is highly correlated with the human-written reference summary gets the highest score and is considered to be the best for the readers.

We have identified two problems with the current evaluation framework. First, the traditional setup may not be suitable for evaluating summaries of user generated text on socio-political issues, primarily because the *implicit bias of the human summarizers* can influence their creation of the reference summaries. We have demonstrated this issue in Section 3 by conducting three surveys - two on the crowdsourcing platform Prolific[1] and one in a classroom setting. We show how summarizers having different opinions can generate very different reference summaries from the same input. *Using such biased reference summaries can lead to wide variation in measures like ROUGE score, depending on who has written the reference summaries.* To our knowledge, no summarization evaluation setup considers this problem. Second limitation of the traditional reference summary-based evaluation, if applied in the context of summarizing user generated content, is that this setup completely ignores the authors whose opinions are being summarized (refer Figure 1).

In this work, we put forward a new perspective: we consider the *summary evaluation of user generated content as a two-sided problem*, and to be fair to both sides, there is a need for metrics that account for both the reader

[1]https://www.prolific.com/

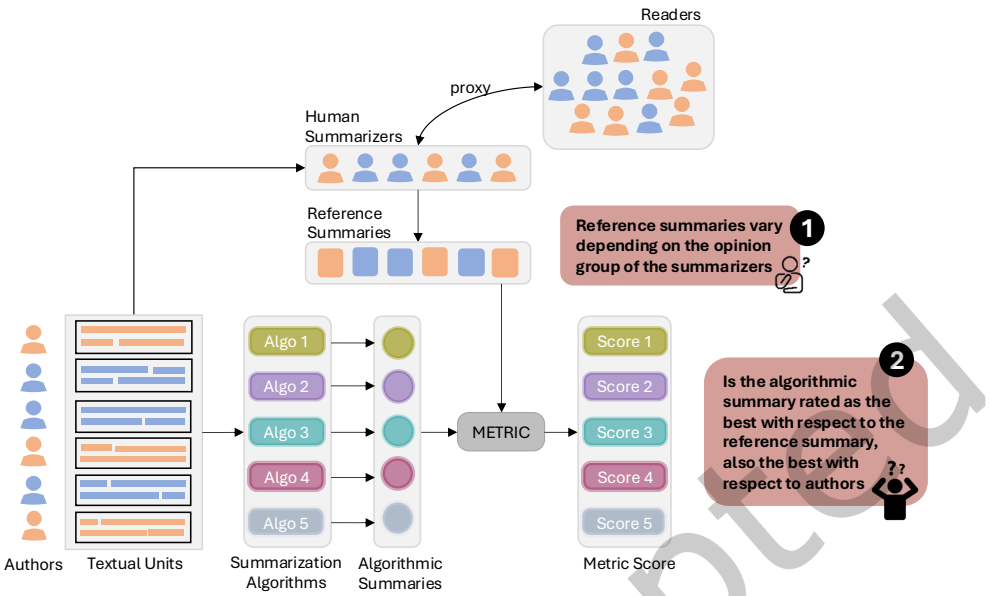


Fig. 1. We identified two major problems with the traditional summary evaluation process. First, the reference summaries vary widely based on the opinion group (,) of human summarizers. Second, the algorithmic summary ranked as the best as per the traditional reference summary based evaluation method, may not be preferred by the authors.

and the author satisfaction. To address the first problem, i.e., implicit bias of summarizers impact the reference summary, we propose an inequality-based variation of reference based metrics (such as ROUGE, BERTScore, MoverScore). This inequality-based variation of metric quantifies *disparities* or *unequal distributions*, i.e., how unevenly the metric score is distributed across different human summarizers. An algorithmic summary with the least inequality score suggests that it effectively represents opinions of different summarizers, making it fairer to the readers. While this work focuses on incorporating inequality into ROUGE, the underlying concept can be extended to other reference summary-based metrics. For the second problem, instead of only relying on reference summaries written by summarizers having unknown biases, we propose a metric that captures the satisfaction of the *authors of the text*. Our motivation stems from the fact that much like readers, authors are also important stakeholders in the summarization process, as it is their opinions that are being summarized. Therefore, it is essential to consider how well the different opinions are represented in the summary. We assume that each author has a preferred set of textual units, and the most suitable summary from the authors' perspective is one that preserves these preferences. We propose an evaluation metric `CROSSEM` (CROwd Satisfaction-based Summary Evaluation Metric) which attempts to capture author satisfaction, and can complement the existing evaluation paradigm.

Contributions: In a nutshell, we make the following contributions in this paper:

• We design a novel survey setup (Section 3) where we first collect authors' opinion on a given topic, and subsequently generate summaries of these responses using both human summarizers and automated summarization algorithms. While we evaluate the algorithmic summaries using the human-written summaries, we gauge author satisfaction by showing the algorithmic summaries to the authors. From the surveys, we observe that (i) inherent

biases of the summarizers result in very different reference summaries for the same input text, and (ii) reference summary based evaluation is inadequate to capture author satisfaction.
• We propose an inequality-based variation of ROUGE, that evaluates an algorithmic summary by assessing how uniformly the ROUGE is distributed across various reference summaries. This proposed variation integrates diverse summarizer perspectives, enabling evaluation of generated summary based on the uniformity with reference summaries.
• We introduce a novel author satisfaction-based evaluation metric CROSSEM, which we found to be highly correlated with the author provided scores.
• To be fair to both readers and authors, we propose a combination of ROUGE and CROSSEM, which can capture both reader and author satisfaction. We conduct sensitivity analysis of this combined metric and observe that it remains stable across a broad range.
• We illustrate how the proposed framework can be applied to social media datasets and demonstrate its effectiveness in capturing the diverse perspectives of both readers and authors.

To our knowledge, ours is the first work to consider summarization of user-generated content as a two-sided problem and the fairness issues therein. While we focus on evaluating summaries in this work, it should spawn future works looking to produce/generate two-sided fair summaries. The remainder of the paper is organized as follows. Section 2 discusses the Background and Related Work. In Section 3, we motivate the need for new evaluation metrics through surveys. Our proposed evaluation metrics are detailed in Section 4 and 5. We apply our proposed metric on a dataset of user generated text from X (formerly Twitter) in Section 6. We conclude the paper in Section 7.

2 Background and Related Work

In this section, we provide a brief overview of text summarization algorithms and their evaluation methods. We also discuss the emerging field of fairness in information retrieval and summarization algorithms, which is closely related to our work.

2.1 Text Summarization Algorithms

With huge amount of data getting generated online, there is a need for algorithms to automatically shorten longer texts and summarize information on a given topic. There are two approaches for automatic summarization - *extractive* and *abstractive*. In extractive summarization, based on the perceived quality and importance, a small subset of the textual units is selected for inclusion in the summary. The aim of an extractive summarization algorithm is to extract a small representative sample of a much larger dataset. On the other hand, abstractive summarization algorithms generate natural language summaries that conveys the most critical information from the original text. In both extractive and abstractive, the goal of the summarizer is to provide the readers a synopsis of textual information.

Over the past years, many text summarization algorithms have been proposed in the literature; the reader can refer to [25, 36] for detailed surveys. Several summarization algorithms meant specifically for user generated data such as microblogs have also been proposed [5, 21, 47, 63, 75, 81, 85, 90, 94]. These approaches primarily attempt to summarize the opinions of people who post on social media [61], encountering new challenges as compared to document summarization. On social media, the text is heterogeneous, individuals express diverse opinions on a given topic, making it crucial for a summary to remain unbiased and not favor any particular group. In contrast, a document is usually authored by a single person, whose opinions remain consistent throughout, i.e., no contradictory perspectives are expected to emerge.

Unsupervised summarization algorithms:
We consider six well-known summarization algorithms. These algorithms generally estimate an importance score

for each textual unit (sentence) in the input, and include the textual units in the summary in decreasing order of their score.
(1) Cluster-rank [30], which clusters the textual units to form a cluster-graph, and uses graph algorithms (e.g., PageRank) to compute the importance of each unit.
(2) DSDR [39], which measures the relationship between the textual units using linear combinations and reconstructions, and generates the summary by minimizing the reconstruction error.
(3) LexRank [26], which computes the importance of textual units using eigenvector centrality on a graph representation based on similarity of the units, where edges are placed depending on the intra-unit cosine similarity.
(4) SumBasic [65], which uses frequency-based selection of textual units, and reweights word probabilities to minimize redundancy.
(5) LSA [34], which constructs a terms-by-units matrix, and estimates the importance of the textual units based on Singular Value Decomposition on the matrix.
(6) LUHN [57], which derives a 'significance factor' for each textual unit based on occurrences and placements of frequent words within the unit.

Supervised summarization algorithms:

With the increasing popularity of neural network based models, the state-of-the-art techniques for summarization have shifted to data-driven supervised algorithms [24].
(7) SummaRuNNer-RNN [64], a Recurrent Neural Network (RNN) based sequence model that provides a binary label to each textual unit – a label of 1 implies that the unit can be part of the summary, while a label of 0 indicates otherwise. Each label has an associated confidence score. The summary is generated by picking textual units labeled 1 in decreasing order of their confidence score, until the desired summary length is exceeded. The proposed model is built around a two-layer bi-directional Gated Recurrent Unit based Recurrent Neural Network (GRU-RNN).[2]
(8) BERTSUM [54] is the SOTA algorithm based on BERT model for extractive summarization.

Large Language Models (LLMs):

In recent years, the advent of LLMs have significantly shifted the NLP paradigm, where instead of fine-tuning the model for a specific task, we have a general-purpose machine that exhibits high performance across various tasks, including summarization [9, 46, 68, 86]. Summaries generated by LLMs showcase high coherence and are overwhelmingly preferred by the human summarizers [55, 73]. We use five LLMs to achieve extractive summaries – **(9)** `llama3-70b-8192` from Meta [87], **(10)** `mixtral-8x7b-32768` from Mistral AI [44], **(11)** `gemini-1.0-pro` from Google [31] and **(12)** `gpt-4o-mini-2024-07-18` from OpenAI [67]. For LLM models we use multi-level prompting strategy [18] to generate extractive summaries by leveraging voting algorithms on top of LLM outputs, which we discuss in the next section.

2.2 Extractive Summarization in LLMs

LLMs inherently generate abstractive summaries by rephrasing the original text. To perform extractive summarization with LLMs, we adopt the multi-level framework LaMSUM introduced in [18]. This method partitions the input into chunks such that each chunk fits within the model's context window. Further, the textual units within each chunk are shuffled to create multiple variants. Each variant is then independently summarized, and the summaries of different variants are aggregated using the proportional voting algorithm [51] to determine the final summary of a given chunk. Summaries from all chunks are then combined to form the input for the next level, and this process is repeated until the desired summary length is achieved. However, in our surveys, the input is

[2]For SummaRuNNer model, the authors of [64] have made the pretrained models available at https://github.com/hpzhao/SummaRuNNer, which are trained on the CNN/Daily Mail news articles corpus.

small enough to be processed in a single chunk, reducing the task to a one-level summarization (Section 3.1). But, for the experiments in Section 6, we employ the full multi-level LaMSUM framework to generate summaries for various social media datasets.

There are several parameters which determine the quality of the LLM output: (i) *Temperature* parameter controls the randomness of the model's output. Low temperature implies more deterministic and repetitive responses, whereas high temperature indicates more creative, diverse less coherent responses; (ii) *Top-p* dynamically chooses a subset of the most probable words, ensuring their combined probability reaches p, rather than considering all possible words. A lower top-p value restricts the model to a smaller set of high-confidence words, while a higher top-p value allows for greater diversity, resulting in more varied outputs; and (iii) *Output tokens* determines the maximum number of tokens the model can generate in a response. A higher token limit produces longer responses and a lower token limit ensures brevity but might cut off important details. Across all experiments with LLMs, we kept the values of Temperature, Top-p and Output tokens as 0, 0.8 and 8192 respectively.

2.3 Evaluation of Summarization Algorithms

Evaluation of algorithmic summaries is challenging, and thus multiple measures have been proposed for such evaluation [84]. Traditionally, the designers of summarization algorithms have relied on human written reference summaries for their evaluation. A human annotator selects the relevant sentences from the original text to create a summary called as reference summary. The summary generated by an algorithm is then evaluated by comparing its similarity with the reference summary; greater the similarity better a summary is considered to be. In our work, we consider some of the most popular metrics for summary evaluation namely ROUGE, BertScore and MoverScore.

(1) ROUGE-L [53] stands for **R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation using **L**ongest Common Subsequence (LCS). LCS is defined as the longest sequence of shared, not necessarily consecutive, ordered words between two sentences. Given a reference summary R and an algorithmic summary $\mathcal{S}$ the union of the longest common subsequences is given by Equation 1. $LCS(r_i, \mathcal{S})$ is the set of longest common subsequences in the sentence r_i from the reference summary and the algorithmic summary $\mathcal{S}$.

$$LCS_{\cup}(R, \mathcal{S}) = \cup_{r_i \in R}\{w \mid w \in LCS(r_i, \mathcal{S})\} \tag{1}$$

ROUGE-L measures the number of overlapping n-gram, word sequences, and word pairs between the algorithmic summary and the human written summary. Equation 2 shows the mathematical formula for recall and precision between algorithmic summary $\mathcal{S}$ and reference summary R. ROUGE-L is an F-score measure as represented in equation 3, parameter β controls the relative importance of the precision and recall.

$$R_{lcs}(R, \mathcal{S}) = \frac{\sum_{r_i \in R} |LCS_{\cup}(r_i, \mathcal{S})|}{numWords(R)} \qquad P_{lcs}(R, \mathcal{S}) = \frac{\sum_{r_i \in R} |LCS_{\cup}(r_i, \mathcal{S})|}{numWords(\mathcal{S})} \tag{2}$$

$$ROUGE - L(R, \mathcal{S}) = \frac{(1 + \beta^2)\, R_{lcs}(R, \mathcal{S})\, P_{lcs}(R, \mathcal{S})}{R_{lcs}(R, \mathcal{S}) + \beta^2\, P_{lcs}(R, \mathcal{S})} \tag{3}$$

(2) BERTScore [93] computes the similarity using BERT-based contextual embeddings between each token present in the algorithmic summary and reference summary. BERTScore has shown high correlation with reader judgements about the goodness of a summary. Recall, precision and F1 score are shown in equation 4 .

$$R_{bert} = \frac{1}{|R|} \sum_{r \in R} max_{s \in \mathcal{S}}(r^T s) \qquad P_{bert} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} max_{r \in R}(r^T s) \qquad F_{bert} = 2\frac{P_{bert} \cdot R_{bert}}{P_{bert} + R_{bert}} \tag{4}$$

(3) MoverScore [96] makes use of Word Mover Distance [50] to measure the semantic distance between an algorithmic summary and a reference summary. In MoverScore, weighted sum of word embeddings are used to

compute n-gram embeddings. For i-th gram, embedding is given by equation 5, where $idf(x_k)$ is the Inverse Document Frequency (IDF) of word x_k computed from all sentences in the corpus and $E(x_k)$ is its word vector. Euclidean distance between the embedding representation of algorithmic summary $\mathcal{S}$ and reference summary R is used to compute the similarity.

$$E(x_i^n) = \sum_{k=i}^{i+n-1} idf(x_k) \cdot E(x_k) \tag{5}$$

All these summary evaluation metrics rely heavily on reference summaries. An algorithmic summary is deemed the best when it closely aligns with the human-written reference summary. The inequality-based approach outlined in Section 4 can be applied across all the metrics discussed above.

2.4 Fairness of Information Filtering Algorithms

To help the users navigate huge amount of online information, all major online platforms today deploy some information filtering algorithms, such as search, summarization or recommendation systems. Given that such information filtering algorithms have far-reaching social and economic consequences in today's world, fairness and anti-discrimination have been recent inclusions in the algorithm design perspective [14, 37]. Several recent works have measured different notions of fairness and biases [3, 8, 14, 40, 71], and attempted to remove the existing unfairness from information filtering algorithms such as clustering [6, 19], ranking [91, 92] and recommendation [15, 69]. Some of them have also focused on two-sided notion of fairness for producers and consumers in recommendation [7, 10, 15, 60, 69, 70, 88, 89] and search systems [22, 28, 29]. In this work, we introduce the idea of two-sided fairness during summarization of user-generated content.

2.5 Fairness in Summarization

Experiments on microblog datasets revealed that existing summarization algorithms often represent socially salient user groups in manner that deviate from their distribution in the original data [80]. Certain groups often receive inadequate representation in the produced summaries, resulting in significantly less visibility compared to what they had in the original data. In order to mitigate these negative effects, several works have attempted to generate high-quality fair summaries [4, 13, 21, 52, 59, 62]. Most of the summarization algorithms use reference summary based metrics like ROUGE for summary evaluation, but such metrics fail to evaluate fairness [95]. Owing to this, concerns have been raised regarding the efficacy of the evaluation approaches [79]. The current work is an attempt to bridge such gaps in the literature.

Prior works have explored summary evaluation in absence of reference summaries, albeit for long documents [11, 12, 23, 56, 74, 76, 78]. While our proposal CROSSEM is driven by similar idea, the motivation is from the fairness perspective. No summary evaluation method takes into consideration the satisfaction of authors who have authored/written the original text. This work proposes an alternative approach to evaluating summaries of user generated text, considering the satisfaction of the readers who read the summary and the authors (members of the crowd) who generate the text. Moreover, a metric focusing on individual satisfaction is a novel contribution in itself.

2.6 Comparison with Earlier Version of this Work:

This article is an extended version of our earlier short paper [17], where we proposed the idea of two-sided fairness in summarization and introduced the CROSSEM metric.

- In our earlier work, we explored non-neural and neural summarization techniques, and found that summaries generated by these models frequently fell short of satisfying the original authors. In recent years, Large Language Models (LLMs) have become the state-of-the-art for a variety of natural language processing tasks,

Topic	Country	No. of Participants	Pre-Screening Question	Opinion Group	Platform
Is Israel's ongoing military operation in **Gaza** justified ?	Worldwide	30	Religious affiliation	Judaism or Islam	Prolific
Why did Kamala Harris lose to Donald Trump in **US election** 2024 ?	US	30	Political affiliation	Democrat or Republican	Prolific
Should there be **reservation** for female candidates in public and private sector jobs in India ?	India	30	Gender	Male or Female	Classroom setting

Table 1. We conducted three surveys on three different topics with participants from various countries. We used pre-screening questions to select participants of different viewpoints, with 50% participants belonging to one opinion group and the remaining 50% belonging to a different opinion group, thus ensuring equal representation of both the groups in a particular survey. A total of 30 participants were selected for each survey.

showing particularly strong performance in summarization [16, 55, 68, 73]. In the current work, we conduct a new set of surveys spanning various topics to assess and compare whether summaries generated by pre-neural methods, neural models and LLMs can achieve author satisfaction.
- We demonstrate that exclusive reliance on reference summary-based evaluation metrics may be unreliable, as the intrinsic biases of summarizers can substantially affect summary quality. To mitigate this limitation, we introduce an inequality-based variant of these metrics.
- In the earlier version of the paper, count vectorizer was employed to represent sentences during the computation of the CROSSEM score. In the current work, we propose utilizing BERT embeddings for sentence representation in the CROSSEM score computation, as they provide a more accurate evaluation of author satisfaction.
- We also introduce inequality based variations of ROUGE and CROSSEM which could ensure individual fairness among readers and authors respectively.
- To establish the generalizability of our findings, we conducted an additional survey involving a larger and more diverse set of participants.
- We demonstrate that a combination of ROUGE and CROSSEM can be used when both reader and author satisfaction are important. Further, we also perform the sensitivity analysis on this combined metric.
- We demonstrate the application of proposed metrics on social media datasets.

Overall, this work builds significantly upon and extends the contributions of our earlier study [17].

3 Pitfalls of traditional summary evaluation: Need for a new paradigm

In this section, we motivate the need for a new approach to evaluate summaries of user generated text, by conducting surveys with participants from three different countries.

3.1 Survey design

We performed three surveys on three debatable topics to get responses from participants having different opinions on them. The first topic for our survey – whether "Israel's ongoing military operation in Gaza is justified?" is crucial due to its humanitarian, legal, and geopolitical implications, affecting global diplomacy, human rights, and regional stability. Different people hold varying opinions – some see it as a necessary act of self-defense against terrorism, while others view it as a disproportionate use of force that worsens the humanitarian crisis and fuels further conflict. The second topic focuses on the 2024 US Election – "Why did Kamala Harris lose to Donald Trump in US election 2024 ?", where people have varying opinions based on their political beliefs. Democrats

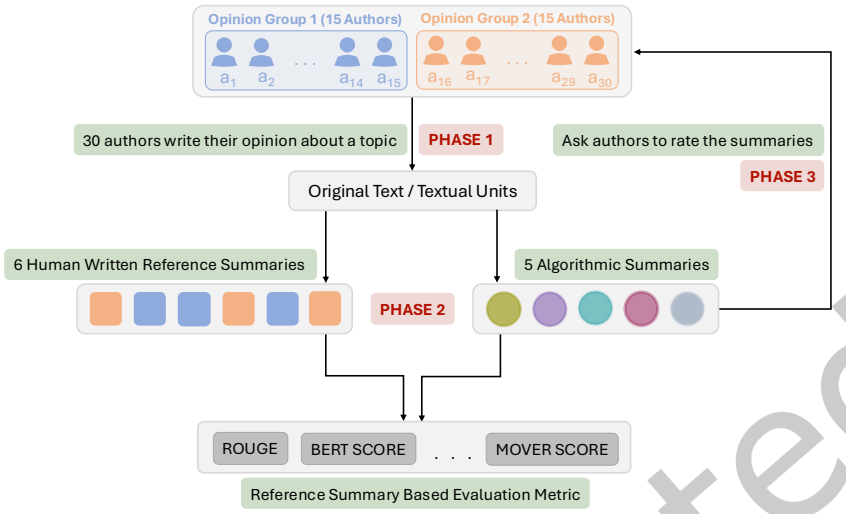


Fig. 2. We conducted each survey in three phases. In the first phase, 30 authors were asked to write their individual opinions about a particular topic. This phase provides us textual units written by different authors (Section 3.1). In the second phase, we ask 6 human summarizers - 3 each from both opinion groups – to summarize the textual units written by 30 authors. We also used 5 automated algorithms to generate extractive summaries. We evaluate the goodness of algorithmic summaries through three reference summary based metrics ROUGE, BERTScore and MoverScore. In the third and final phase, we go back to the authors to know their satisfaction with each algorithmic summary (Section 3.3).

attribute Kamala Harris's defeat to inadequate campaign planning and unfavorable media coverage, whereas Republicans interpret it as a clear rejection of progressive policies by the electorate. Third topic addresses the issue of gender equality, workforce diversity, and historical disadvantages faced by women in India – "Should there be reservation for female candidates in public and private sector jobs in India ?". It sparks debate on whether reservations or alternative policies are the best way to ensure fair representation in public and private sector jobs. Table 1 presents the details about the surveys conducted with participants from different countries of the world. We used Prolific[3] for surveys on `Gaza` and `US-Election`. During the registration on Prolific, users are required to provide responses to a set of screening questions, thereby establishing prior knowledge of the user's opinion group [72]. The participants for each survey were selected based on their answer to the pre-screening question (as detailed in Table 1). For every survey, the first 15 respondents from each opinion group who answered the survey question were chosen as the participants for our study. Consequently, for `Gaza` survey, 15 participants are Jews, while another 15 are Muslims. In `US-Election` survey, 15 participants are democrats, and other 15 are republicans. For the survey on `Reservation`, we could not use Prolific as the number of available respondents from India were not adequate for our survey. Hence, we conducted the survey in a classroom setting with final year students from an Indian University, as the topic is very relevant to students as they prepare to enter the job market after completing their studies. To get the perceptions of different gender groups, 15 male and 15 female students were chosen to be a part of the survey.

For all three surveys, we followed the same course of action. Each survey had three different phases; a diagrammatic view of the same is shown in Figure 2. In the **first phase**, 30 participants (we call them *authors*) were asked to write 4-5 sentences about their respective topic. For instance, in the topic related to `Gaza`, 30 authors

[3]https://www.prolific.co/

Method	Avg All	Rank All	Avg Jews	Rank Jews	Avg Muslim	Rank Muslim
ROUGE						
LexRank	0.409	3	0.366	5	0.494	1
LSA	0.391	4	0.375	4	0.406	2
BERTSUM	0.349	5	0.378	3	0.320	5
Llama3-70b	0.415	2	0.467	1	0.337	4
GPT-4o-mini	0.423	1	0.453	2	0.364	3
BERTScore						
LexRank	0.809	5	0.804	5	0.818	2
LSA	0.825	2	0.83	3	0.820	1
BERTSUM	0.813	4	0.824	4	0.802	4
Llama3-70b	0.820	3	0.837	2	0.796	5
GPT-4o-mini	0.838	1	0.85	1	0.816	3
MoverScore						
LexRank	0.579	5	0.575	5	0.588	1
LSA	0.587	2	0.591	3	0.583	2
BERTSUM	0.581	4	0.589	4	0.572	4
Llama3-70b	0.587	2	0.601	2	0.566	5
GPT-4o-mini	0.596	1	0.604	1	0.581	3

Table 2. Ranking of different summarization algorithms based on ROUGE, BERTScore and MoverScore for the Gaza survey. Avg All is the average metric score when all 6 reference summaries are considered. Rank All is the ranking obtained on the basis of Avg All. Avg Jews is the average metric score obtained by considering only the reference summaries written by people whose religious affiliation is Judaism. Similarly, Avg Muslim is the average metric score obtained by considering only the reference summaries of 3 summarizers whose religion is Islam. Rank Jews and Rank Muslim are the rankings obtained on the basis of Avg Jews and Avg Muslim respectively.

wrote their opinions about *"Is Israel's ongoing military operation in Gaza justified ?"*. Likewise, for the survey on `US-Election`, 30 US-based authors put their thoughts on *"Why did Kamala Harris lose to Donald Trump in US Election 2024 ?"*, and for the topic related to `Reservation`, 30 Indian authors wrote their views about *"Should there be reservation for female candidates in government and private sector jobs in India ?"*.

From Phase 1, we received a set of textual units (original text written by the authors) for a particular topic. We then apply 5 summarization algorithms described in Section 2.1 on the textual units to obtain 5 different algorithmic summaries.[4] The prevailing method for measuring the goodness of algorithmic summaries is to *compare them with human written reference summaries*. Thus, in the **second phase** of the survey, we ask 6 additional participants to select the most relevant 5 textual units[5] which summarize the texts written by the authors. We call these participants as *human summarizers*. We select them based on their leaning observed in the pre-screening question. Out of these 6 summarizers, 3 belong to one opinion group and other 3 belong to the other opinion group. Three different evaluation metrics dependent on reference summaries – ROUGE, BERTScore and MoverScore (introduced in section 2.3) are utilized to calculate the goodness of the algorithmic summaries. Note that in this work we have used only three metrics, but the same approach can be easily extended to any evaluation metric dependent on reference summaries.

[4]We limit our selection to only 5 summarization algorithms from Section 2.1, as presenting 12 different summaries could potentially overwhelm the authors, making it difficult for them to accurately rank the outputs.

[5]In this paper, our focus is on extractive summarization, and hence the summarizers are asked to select 5 textual units from 30 textual units written by 30 different authors.

Method	Avg All	Rank All	Avg Democrat	Rank Democrat	Avg Republican	Rank Republican
ROUGE						
LexRank	0.296	3	0.482	1	0.203	5
LSA	0.314	2	0.295	2	0.332	2
BERTSUM	0.278	5	0.243	5	0.312	4
Llama3-70b	0.376	1	0.29	3	0.434	1
GPT-4o-mini	0.286	4	0.259	4	0.314	3
BERTScore						
LexRank	0.779	5	0.789	5	0.774	5
LSA	0.816	2	0.806	1	0.826	2
BERTSUM	0.806	3	0.793	2	0.819	3
Llama3-70b	0.818	1	0.791	3	0.837	1
GPT-4o-mini	0.804	4	0.791	3	0.818	4
MoverScore						
LexRank	0.556	5	0.567	2	0.55	5
LSA	0.575	2	0.570	1	0.58	2
BERTSUM	0.568	3	0.561	4	0.575	4
Llama3-70b	0.587	1	0.564	3	0.602	1
GPT-4o-mini	0.568	3	0.56	5	0.576	3

Table 3. Ranking of different summarization algorithms based on ROUGE, BERTScore and MoverScore for the `US-Election` survey. Avg All is the average metric score when all 6 reference summaries are considered. Rank All is the ranking provided to summarization algorithm on the basis of Avg All. Avg Democrat is the average metric score obtained by considering the summaries of democratic summarizers. Similarly, Avg Republican is the average metric score obtained by considering the summaries of republican summarizers. Rank Democrat and Rank Republican is the ranking obtained on the basis of Avg Democrat and Avg Republican respectively.

Finally, in the **third phase**, we go back to the authors and ask them to rate the algorithmic summaries themselves in the scale of [1 – 10]. The details can be found in Section 3.3.

3.2 How summarizers opinion influence reference summaries ?

As mentioned earlier, the traditional and prevalent way of evaluating a summarization algorithm is to compare the algorithmically generated summary with one or multiple human-generated reference summaries. Tables 2, 3 and 4 show ROUGE, BERTScore and MoverScore between different algorithmic and reference summaries for `Gaza`, `US-Election` and `Reservation` respectively. In all three tables, *Avg All* refers to the average metric score when we consider reference summaries from all six human summarizers. Similarly, *Rank All* is the ranking of the summarization algorithms on the basis of *Avg All*. The other scores and rankings are obtained based on reference summaries written by summarizers with particular opinions/leanings. For example, in Table 3, *Avg Democrat* is the average metric score obtained by considering the reference summaries of 3 summarizers who have political affiliation as Democrat. *Rank Democrat* is the ranking provided on the basis of *Avg Democrat*. On the other hand, *Avg Republican* is the average metric score obtained by considering the reference summaries of 3 summarizers whose political affiliation is Republican. *Rank Republican* is the ranking provided on the basis of *Avg Republican*.

A better rank for a summarization algorithm is indicative of its acceptance over others for the summarization task. For instance, in Table 3 for ROUGE metric, if the reference summaries only come from the human summarizers whose political leaning is Democrat (*Rank Democrat*), LexRank would be the algorithm of choice; whereas when considering only the summaries of people whose religion is Republican (*Rank Republican*), Llama emerges as

Method	Avg All	Rank All	Avg Female	Rank Female	Avg Male	Rank Male
ROUGE						
LexRank	0.344	5	0.333	4	0.355	5
LSA	0.395	3	0.365	2	0.425	2
BERTSUM	0.447	1	0.504	1	0.389	4
Llama3-70b	0.372	4	0.305	5	0.394	3
GPT-4o-mini	0.397	2	0.34	3	0.453	1
BERTScore						
LexRank	0.817	5	0.819	4	0.815	5
LSA	0.837	2	0.837	2	0.837	2
BERTSUM	0.851	1	0.867	1	0.835	3
Llama3-70b	0.822	4	0.817	5	0.823	4
GPT-4o-mini	0.834	3	0.826	3	0.842	1
MoverScore						
LexRank	0.582	5	0.583	3	0.581	5
LSA	0.599	2	0.601	2	0.598	1
BERTSUM	0.603	1	0.614	1	0.591	3
Llama3-70b	0.586	4	0.581	5	0.587	4
GPT-4o-mini	0.589	3	0.583	3	0.596	2

Table 4. Ranking of different summarization algorithms based on ROUGE, BERTScore and MoverScore for the Reservation survey. Avg All is the average metric score when all 6 reference summaries of Male and Female students are considered. Rank All is the ranking provided to summarization algorithm on the basis of Avg All. Avg Female is the average metric score obtained by considering only the Female written summaries. Avg Male is the average metric score obtained by considering the Male written summaries. Rank Female and Rank Male are the rankings obtained on the basis of Avg Female and Avg Male respectively.

the top-ranked algorithm. Essentially, for US-Election survey, the ranking along *Rank Democrat* and *Rank Republican* columns are different across all metrics. Similar trends can be observed for Gaza and Reservation surveys in Tables 2 and 4 as well. In Gaza survey, for ROUGE metric, spearman [82] and kendall-tau [49] correlation coefficients between Rank Jews and Rank Muslim are -0.7 and -0.6 respectively, indicating that summarizers may produce completely different summaries due to their inherent biases.

Through these surveys we can conclude that the reference summaries of crowd-sourced text can vary significantly, based on the inherent ideological biases/leanings of the summarizers who write the reference summaries (refer Figure 1). Such biases of human summarizers are not considered in the traditional evaluation approach such as ROUGE.

3.3 Going back to the authors

Now, we explore another potential limitation of the traditional way of evaluating algorithmic summarization, which is based on comparing algorithmic summaries with the reference summaries. The underlying motivation behind including the reference summaries in the evaluation process is to assess how well an algorithm-generated summary matches a human-written summary. Human summarizers who write the reference summaries act as a proxy for the readers, i.e., they try to create a summary that will eventually satisfy the readers. In summarization of user generated texts (such as, tweets or Facebook posts), alongside the readers, the authors who wrote the input text are also important stakeholders, since it is their opinions which are being summarized. Identifying the stakeholders is especially relevant for polarizing topics, since authors of different viewpoints would like their different opinions to be reflected in the summary. Thus, rather than relying only on human summarizers with

Method	Gaza		US-Election		Reservation	
	Author Rating	Rank	Author Rating	Rank	Author Rating	Rank
LexRank	6.263	2	6.682	1	7.600	3
LSA	5.895	4	6.409	2	7.800	1
BERTSUM	6.000	3	6.045	4	6.267	5
Llama3-70b	6.316	1	5.909	5	7.733	2
GPT-4o-mini	5.632	5	6.409	2	7.200	4

Table 5. Rating and rank provided by authors for Gaza, US-Election and Reservation survey. Authors provide rating to each algorithmic summary on a scale of 10. Values shown in Author Rating column are average over 30 authors. Rank is the ranking obtained on the basis of Author Rating.

unknown implicit biases, it might be a good idea to go back to the authors to see how they perceive different algorithmic summaries.

In the third phase of each survey, we showed 5 algorithmic summaries back to the 30 authors, and asked them to rate each algorithmic summary on a scale of [1 – 10]. Higher score is indicative of higher satisfaction with the summary. Table 5 shows the average rating provided by the authors to various algorithmic summaries and resultant ranking of the algorithms. For all three surveys, it can be observed that *Avg All* rankings in Tables 2, 3, 4 are different from the rankings computed from the average author ratings in Table 5. For instance, if we observe *Avg All* column in Table 2, GPT has the highest ROUGE score, BERTScore and MoverScore; however, the authors seem to be most satisfied with the summary produced by Llama (Table 5). Similarly, in Table 3 for US-Election survey, Llama gets the highest ROUGE, BERTScore and MoverScore, but authors are happier with the summaries provided by LexRank. In Table 4 for Reservation survey, BERTSUM algorithm has the highest score for all the metrics, whereas the authors are most satisfied with the summary generated by LSA.

From these observations, we can conclude that the traditional summary evaluation metrics dependent on human written reference summaries are not able to capture author satisfaction (Figure 1). At the same time, it is not possible for the authors to explicitly provide their preferences over different algorithmic summaries. Hence, we need an automated approach to capture author satisfaction, which we discuss in the next section.

3.4 Takeaway

Through surveys, we uncovered two key shortcomings of traditional summary evaluation methods.

Pitfall 1 (PF1): Human summarizers may carry implicit biases, resulting in generation of biased reference summaries. Evaluation metrics that rely on reference summaries, such as ROUGE, can lead to the selection of biased algorithmic summaries due to biases present in reference summaries. Consequently, selection of biased algorithmic summary might lead to the over-representation of opinions from one group, marginalizing or excluding other groups.

Pitfall 2 (PF2): Traditional summary evaluation methods, such as ROUGE, prioritize selecting summaries that maximize reader satisfaction. However, authors are also key stakeholders, and it is crucial to ensure that diverse author perspectives are represented in the summary. However, no existing evaluation methods take into account the author satisfaction.

4 Addressing PF1: Inequality-Driven Metric to Account for Summarizer Bias

Traditional summary evaluation method which involves human written summaries (such as ROUGE, BERTScore and MoverScore), aims to capture reader satisfaction. *The human summarizers are a proxy for the readers* and their target is to write a summary that can satisfy the readers or make the readers happy. But the human summarizers might have implicit bias and they may write a summary which favors a particular opinion group [27, 32, 43].

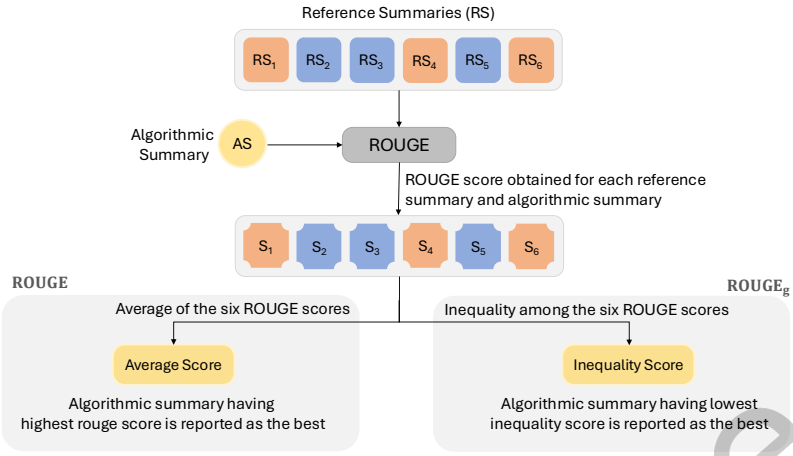


Fig. 3. In our setup, we obtain an algorithmic summary and six reference summaries written by human summarizers. ROUGE captures the similarity between an algorithmic summary and each reference summary, yielding six scores. In traditional setup, for calculation of the final ROUGE score, an average is taken over the six ROUGE scores. Whereas, in our proposed metric ROUGE_g, the final score is based on the inequality among the six ROUGE values.

Consider example given in Figure 4, let a_1, a_2, a_3, a_4 are male authors belonging to opinion group 1 and a_5, a_6, a_7, a_8 are female authors belonging to opinion group 2. AS_1 is a biased algorithmic summary as it includes textual units only from the male authors, whereas AS_2 is an unbiased summary that represents viewpoints from both male and female groups in equal proportion. Reference summaries RS_1, RS_3, and RS_4 are authored by male summarizers, while RS_2 is created by a female summarizer. Given the predominance of male summarizers, there is an imbalance in the reference summaries, with a greater number reflecting male-favoring opinions and fewer incorporating female-favoring perspectives. Evaluation metrics (such as ROUGE) are used to assess the quality of algorithmic summaries by measuring their overlap with the reference summaries. The summary with the highest overlap score is deemed the most effective. Due to the disproportionate representation of male viewpoints in the reference summaries, a biased summary like AS_1, which aligns with male opinions, will naturally achieve a higher ROUGE score. This introduces a significant concern: if the reference summaries themselves are biased, then ROUGE is likely to favor and rank biased algorithmic summaries more highly. This observation underscores a fundamental challenge in summary evaluation – biases in human-written references, often implicit and unconscious, can distort the evaluation process. It is inherently difficult to eliminate such biases entirely because they are deeply rooted in individual perspectives and societal structures. This leads to an important question: in the presence of biased reference summaries, how can we ensure fair and objective evaluation of algorithmic summaries? Addressing this issue is crucial for the development of more equitable and accurate automatic summarization systems.

One way to address this problem is to carefully select the summarizers who write the reference summaries, such as, ensuring balanced representation of annotators from different opinion groups. The reference summaries should be written either by summarizers who are having a neutral viewpoint, or by a set of summarizers who have equal participation from male and female groups. However, it is difficult to implement this approach in practice, since biases of human summarizers are often latent and implicit, and can not be known in advance. Additionally, the complexity of global social dynamics adds another layer of uncertainty. Sociopolitical contexts vary widely across regions and cultures, and the ways in which biases manifest can differ depending on the

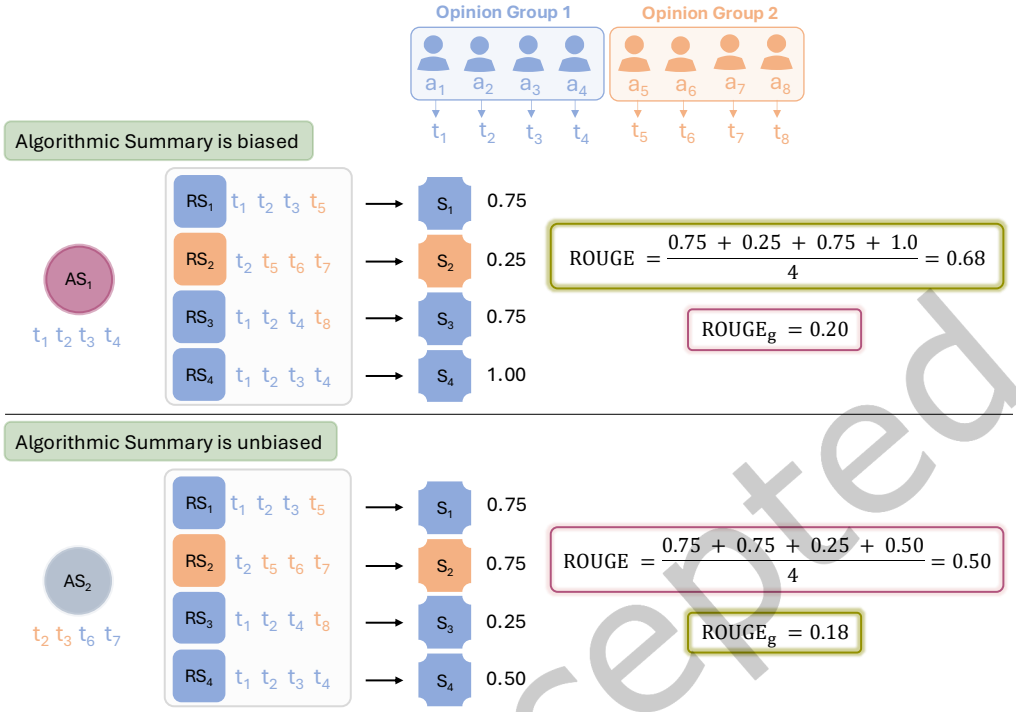


Fig. 4. Example illustrating the utility of ROUGE_g over ROUGE. Authors a_1, a_2, a_3, and a_4 are male authors belonging to opinion group 1 and author textual units t_1, t_2, t_3, and t_4 respectively. Authors a_5, a_6, a_7, and a_8 are female authors and belong to opinion group 2 and write textual units t_5, t_6, t_7 and t_8 respectively. S_i represents the ROUGE score between algorithmic summary AS and RS_i. We examine two scenarios: one where the algorithmic summary AS_1 reflects a biased perspective and another where AS_2 is unbiased. Based on the ROUGE score, AS_1 would be favored over AS_2. ROUGE evaluates overlap between the algorithmic summary and the reference summaries, and since the reference summaries are biased towards male, thus ROUGE inadvertently rewards algorithmic summary that is biased towards male. In contrast, ROUGE_g prioritizes an algorithmic summary that show uniform overlap across the reference summaries. According to ROUGE_g, AS_2 is favored, because it includes textual units from multiple opinion groups, resulting in a more uniform overlap with the reference summaries.

issue at hand. This makes it even more difficult to predict how a summarizer's background or viewpoint might influence their summary, further complicating efforts to ensure fairness and objectivity.

A second approach to addressing the issue of bias from human summarizers involves using an inequality-based evaluation method. In traditional reference summary-based metrics like ROUGE, an algorithmic summary $\mathcal{S}$ is evaluated by comparing it to multiple human-written reference summaries R_i. The final ROUGE score is computed as the arithmetic mean of these individual similarity scores, as described in Equation 3. However, the mean alone does not capture how evenly the summary aligns with the different reference summaries — meaning it overlooks any imbalance or inequality in the score distribution across them. To promote fairness in summary evaluation, we propose an inequality based variant of the ROUGE metric. The core idea is to prefer algorithmic summaries that exhibit a uniform level of semantic overlap with all the reference summaries. In other words, a fair summary should not align heavily with just a subset of the references (e.g., those written by a particular

opinion group), but instead show consistent agreement across all references. Such a summary will result in a more even distribution of similarity scores and therefore exhibit lower inequality, as illustrated in Figure 3 and 4.

To quantify this inequality, we use the Gini coefficient, a statistical measure originally developed to assess income inequality [33]. Gini coefficient is based on the difference between the Lorenz curve and a perfectly equal distribution line. A Gini coefficient of 0 indicates complete equality, meaning the similarity scores are uniformly distributed across all references. On the other hand, a Gini coefficient of 1 (or 100%) reflects extreme inequality, where one reference summary dominates the similarity, while others contribute little or none. Applying this concept to summary evaluation allows us to identify algorithmic summaries that are equitably aligned with diverse reference summaries, thus offering a way to mitigate the influence of biased reference summaries.

Consider a system evaluated against multiple reference summaries – where the similarity scores in two scenarios are {0.5, 0.4, 0.5, 0.4} (Case 1) and {0.8, 0.8, 0.1, 0.1} (Case 2). Although both cases yield the same mean value, but they reflect different pattern. Case 1 corresponds to a summary that aligns uniformly with all reference summaries, indicating that it captures elements from each of them. In contrast, case 2 represents a biased summary that aligns strongly with only a subset of references while largely ignoring the others. Case 1 arises when the algorithmic summary has some overlap with each of the reference summaries – some textual units from each reference summary is present in the algorithmic summary. On the contrary case 2 occurs when the algorithmic summary has very high overlap with some reference summaries and little to none with other reference summaries. From the fairness perspective, case 1 is preferable, as it treats all human summarizers who serve as proxies for the readers more equitably. While ROUGE assigns similar evaluations to both cases due to identical mean scores, the Gini coefficient distinguishes between them by favoring distributions with greater uniformity. The Gini coefficient ensures fairness by moving evaluation from *"how much overlap exists"* to *"how evenly the overlap is distributed"*. Consequently, Equation 3 can be modified to account for individual biases of the summarizers:

$$\text{ROUGE}_\text{g}(\mathcal{S}) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}|ROUGE(R_i,\mathcal{S}) - ROUGE(R_j,\mathcal{S})|}{2\cdot K\cdot\sum_{j=1}^{K}ROUGE(R_j,\mathcal{S})} \tag{6}$$

where K is the number of summarizers and $ROUGE(R_i,\mathcal{S})$ is ROUGE score between reference summary R_i and algorithmic summary $\mathcal{S}$. ROUGE_g defines the inequality in the ROUGE score. ROUGE_g value close to 0 suggests that a given algorithmic summary has equal overlapping with each of the reference summaries, indicating minimal inequality. In contrast, a score close to 1 signifies a high level of inequality i.e., some reference summaries are close to algorithmic summary and some reference summaries are very distant.

5 Addressing PF2: Author Satisfaction based Metric for Summary Evaluation

In Section 3.3, we saw that the ranking of algorithms based on author scores does not match the ranking based on ROUGE or other reference-summary based evaluation metrics. However, it is impractical to ask the authers to rate the quality of summary for every topic. Hence, in this section, we propose a family of metrics named CROSSEM (**CRO**wd **S**atisfaction-based **S**ummary **E**valuation **M**etric) which can automatically capture the satisfaction of authors and thus can be used for summary evaluation tasks.

5.1 Defining CROSSEM

The underlying idea behind CROSSEM is that a good summary of user generated text should satisfy the members of the crowd who wrote the text (the authors). We intuitively assume that an author will be satisfied if his/her opinion (as expressed in the textual unit written by him/her) is included in the summary. Thus, a good summary should try to maximize the satisfaction of the authors who wrote the textual units that are being summarized. In other words, we propose to view a summarization approach similar to an *election process* in a democratic society. Here we consider the textual units to be the candidates in the election, the authors are the voters, and

the summary is the set of elected candidates. We assume that each author (voter) has a set of *approved textual units (candidates) that he/she would like to be included in the summary*; this is as if an author votes for a subset of the textual units. We have a notion of 'satisfaction' of an author (discussed below), and we propose to evaluate a summary based on how it captures the satisfaction of all authors. With this underlying rationale, we now formally define the author satisfaction-based summary evaluation.

Definition 1 (CROwd Satisfaction-based Summary Evaluation Metric). *Let $\mathcal{T} = \{t_1, t_2, \dots t_N\}$ be a set (universe) of textual units, written by a set of authors $\mathcal{A} = \{a_1, a_2, \dots a_N\}$, where $t_i \in \mathcal{T}$ has been written by $a_i \in \mathcal{A}$. We consider a textual unit as a single post. Each author a_i has an approved subset of textual units $V_i \subseteq \mathcal{T}$ which he/she would like to see in the summary.* [6] *Every summarization algorithm has as input $\mathcal{T}$ and an integer $\mathcal{B}$ (budget) which denotes the maximum number of textual units that a summary can contain, and it outputs a summary $\mathcal{S} \subseteq \mathcal{T}$ with $|\mathcal{S}| \leq \mathcal{B}$. The summary $\mathcal{S}$ would be evaluated based on $\sum_{i=1}^{N} sat(a_i, \mathcal{S})$ where $sat(a_i, \mathcal{S})$ is a measure of how satisfied the author a_i is with the summary $\mathcal{S}$.* CROSSEM *measures the mean (average) satisfaction of all authors for a given summary.*

$$\texttt{CROSSEM}(\mathcal{S}) = \frac{1}{N} \sum_{i=1}^{N} sat(a_i, \mathcal{S}) \tag{7}$$

In other words, the quality of a summary $\mathcal{S}$ will be measured based on how well the summary optimizes the satisfaction of all the authors. To estimate the satisfaction of an author a_i with the generated summary $\mathcal{S}$, we evaluate the extent to which the author's ideas are captured in $\mathcal{S}$. This is achieved by computing the semantic similarity between V_i and $\mathcal{S}$.

$$sat(a_i, \mathcal{S}) = sim(V_i, \mathcal{S}) \tag{8}$$

where $sim(V_i, \mathcal{S}) \in [0, 1]$ is a measure of the semantic similarity between two sets of textual units. A higher similarity score indicates a stronger alignment between the author's original ideas and the generated summary, and thus reflect greater author satisfaction. An author is considered satisfied when their ideas or similar ideas are represented in the summary.

The semantic similarity $sim(V_i, \mathcal{S})$ can be computed using a variety of approaches. There is a long line of research on measuring semantic similarity between two pieces of text, ranging from TF-IDF similarity, to using word embeddings and neural models. In the earlier version of this work [17], we employed a count vectorizer to transform the text into a vector. Count vectorizer tokenizes the text data and assigns a vector on the basis of the frequency (count) of each word. Accordingly, each textual unit in the approved set V_i and in the summary $\mathcal{S}$ is converted into its corresponding vector form. Equation 9 represents the the computation of semantic similarity $sim(V_i, \mathcal{S})$ using count vectorizer. Here, $|V_i|$ denotes the number of textual units in the approved set V_i authored by a_i, and $|\mathcal{S}|$ shows the number of textual units present in the summary $\mathcal{S}$. $CV(V_i^j)$ and $CV(\mathcal{S}^k)$ correspond to the count vectorizer representation of the j^{th} and k^{th} textual unit in V_i and $\mathcal{S}$ respectively. Cosine similarity is used to calculate the pairwise similarity between these vectors. The pairwise cosine similarity between vectors of V_i and $\mathcal{S}$ is then aggregated. To ensure the similarity score lies within the range [0,1], the aggregate value is normalized by the total number of textual units in V_i and $\mathcal{S}$.

$$sim(V_i, \mathcal{S}) = \frac{1}{|V_i|\,|\mathcal{S}|} \sum_{j=1}^{|V_i|} \sum_{k=1}^{|\mathcal{S}|} cosine(CV(V_i^j), CV(\mathcal{S}^k)) \tag{9}$$

[6] The set V_i comprises of the textual units written by the author a_i.

Method	Gaza		US-Election		Reservation	
	CROSSEM Score	Rank	CROSSEM Score	Rank	CROSSEM Score	Rank
LexRank	0.715	3	0.787	1	0.691	3
LSA	0.708	4	0.760	4	0.715	1
BERTSUM	0.731	2	0.758	5	0.666	4
Llama3-70b	0.743	1	0.769	2	0.700	2
GPT-4o-mini	0.688	5	0.768	3	0.652	5

Table 6. CROSSEM score and rank achieved by summarization algorithms for Gaza, US-Election and Reservation survey.

Sentence representation can have a huge impact on the task performance [48]. We also experiment with distilbert-base BERT embeddings. Textual unit V_i^j and $\mathcal{S}^k$ are represented through BERT embeddings as shown in Equation 10. Thereafter, cosine similarity is used to gauge the semantic closeness between BERT embeddings of V_i^j and $\mathcal{S}^k$. Our proposed $sim(V_i, \mathcal{S})$ function can very well handle the case when multiple posts are authored by a user i.e., $|V_i| > 1$.

$$sim(V_i, \mathcal{S}) = \frac{1}{|V_i|\,|\mathcal{S}|} \sum_{j=1}^{|V_i|} \sum_{k=1}^{|\mathcal{S}|} cosine(BERT(V_i^j), BERT(\mathcal{S}^k)) \tag{10}$$

We experimented with two different sentence representations – count vectorizer and BERT embeddings, and found that usage of BERT embeddings during computation of $sim(V_i, \mathcal{S})$ revealed better results.[7] In the subsequent sections, for our survey experiments, we report similarity scores using BERT embeddings as it gives more superior results. CROSSEM is defined to have values in the range $[0.0, 1.0]$ where a higher value indicates greater degree of satisfaction.

Two interesting points can be noted about CROSSEM:
(1) These metrics can be applied in the absence of any reference summaries, thus can be continuously observed in a real-world summarization setting.
(2) The metric CROSSEM can be applied to evaluate both *extractive and abstractive summaries* by defining the similarity function *sim* suitably. However, in the present work, we are only considering extractive summarization.

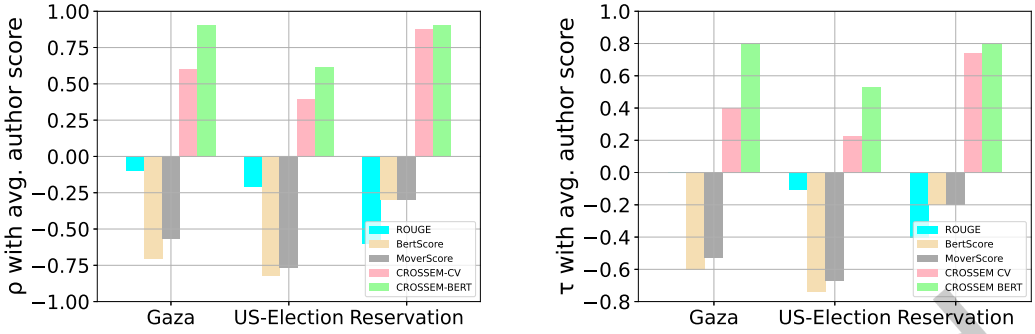
5.2 Can CROSSEM capture author satisfaction ?

We check the performance of CROSSEM on the data obtained from three surveys. Table 6 shows the result of CROSSEM computed based on the algorithmic summaries and textual inputs obtained through Gaza, US-Election and Reservation surveys. It can be observed that the rankings obtained by CROSSEM (Table 6) are close to the rankings obtained from average author ratings (Table 5). For Gaza, US-Election and Reservation survey, both CROSSEM and author scores reveal the best algorithms to be Llama, LexRank and LSA respectively. To quantify the closeness in ranking through different metrics, we use two rank based correlation metrics:
(1) Spearman Rank Correlation [82] assesses the strength and direction of association between two ranked variables by defining a monotonic function. Formally, Spearman correlation coefficient ρ is computed as:

$$\rho = \frac{cov(R(x), R(y))}{\sigma_{R(x)} \cdot \sigma_{R(y)}} \tag{11}$$

Here, ρ denotes Spearman correlation coefficient, $cov(R(x), R(y))$ is covariance between the rank of variables x and y. $\sigma_{R(x)} \cdot \sigma_{R(y)}$ denotes the standard deviation of rank variables.

[7] LLMs can also be employed to estimate semantic similarity, which we plan to explore in future work.



(a) Spearman Correlation

(b) Kendall Tau Correlation

Fig. 5. (a) Spearman's ρ and (b) Kendall's τ correlation coefficients computed between different summary evaluation measures on Gaza, US-Election and Reservation surveys. ROUGE, BERTScore and MoverScore are negatively correlated with average author ratings; whereas CROSSEM has positive correlation with the average author ratings across all three surveys.

(2) Kendall Rank Correlation [49] measures the association between two variables based on the number of concordances and discordances in paired observations. Two observations (X_i, Y_i) and (X_j, Y_j) are concordant if they appear in relative same order i.e., if $X_i < X_j$ then $Y_i < Y_j$ OR if $X_i > X_j$ then $Y_i > Y_j$. Discordant pairs are the ones who are not concordant. Kendall's τ is calculated as:

$$\tau = \frac{\text{no. of concordant pairs - no. of discordant pairs}}{\text{no. of pairs}} = 1 - \frac{2 \times \text{no. of concordant pairs}}{\binom{n}{2}} \tag{12}$$

Figures 5(a) and 5(b) show the rank correlation of average author ratings with the summary evaluation metrics ROUGE, BERTScore, MoverScore, and two variants of the proposed CROSSEM metric — one using Count Vectorizer and the other using BERT embeddings. These correlations demonstrate how well each metric aligns with the satisfaction of the authors. The results reveal a noteworthy trend: the traditional reference based metrics – ROUGE, BERTScore and MoverScore, exhibit a negative correlation with the average author ratings. This suggests that the metrics which rely on comparing the generated summaries to reference summaries, often fail to reflect what authors actually value in a good summary. In other words, even if a summary scores highly according to these traditional metrics, it may still be judged poorly by the authors, indicating a mismatch between metric outputs and authors perception of quality. In contrast, the proposed CROSSEM metric demonstrates a consistent positive correlation with the author ratings across all three user surveys analyzed. This indicates that CROSSEM better aligns with authors and could serve as a more reliable proxy for measuring author satisfaction with summaries. Notably, when comparing the two variants of CROSSEM, the version that uses BERT embeddings shows a higher correlation with author scores than the one based on the Count Vectorizer. This implies that BERT embeddings capture deeper contextual and semantic information from the text and are more effective in modeling the aspects of summaries that authors consider important. Overall, these findings suggest that reference-free evaluation methods like CROSSEM, especially when leveraging powerful embedding models like BERT, are more capable of assessing summary quality in a way that resonates with authors preferences.

5.3 Generalization

To validate the generalizability of our findings, we conducted an additional experiment on Prolific with a larger and more diverse participant pool, following a survey design similar to that outlined in Section 3.1. A total of 200 participants – 100 self-identifying as members of the LGBTQ+ community and 100 as non-LGBTQ+ – were

Method	Avg All	Rank All	Avg LGBTQ+	Rank LGBTQ+	Avg Non LGBTQ+	Rank Non LGBTQ+
ROUGE						
LexRank	0.555	3	0.647	2	0.463	4
LSA	0.530	4	0.462	5	0.597	1
BERTSUM	0.556	2	0.574	3	0.537	2
Llama3-70b	0.561	1	0.677	1	0.444	5
GPT-4o-mini	0.505	5	0.546	4	0.464	3
BERTScore						
LexRank	0.802	5	0.799	4	0.805	4
LSA	0.833	4	0.798	5	0.867	1
BERTSUM	0.865	1	0.923	1	0.806	3
Llama3-70b	0.861	2	0.917	2	0.805	4
GPT-4o-mini	0.857	3	0.905	3	0.808	2
MoverScore						
LexRank	0.568	5	0.566	5	0.569	5
LSA	0.604	4	0.573	4	0.635	1
BERTSUM	0.644	1	0.712	1	0.576	2
Llama3-70b	0.641	2	0.705	2	0.576	2
GPT-4o-mini	0.635	3	0.694	3	0.575	4

Table 7. Ranking of different summarization algorithms based on ROUGE, BERTScore and MoverScore for the LGBTQ+ survey. Avg All is the average metric score when all 30 reference summaries are considered. Rank All is the ranking obtained on the basis of Avg All. Avg LGBTQ+ is the average metric score obtained by considering only the reference summaries written by summarizers of LGBTQ+ community. Similarly, Avg Non LGBTQ+ is the average metric score obtained by considering only the reference summaries of 15 summarizers who do not belong to LGBTQ+. Rank LGBTQ+ and Rank Non LGBTQ+ are the rankings obtained on the basis of Avg LGBTQ+ and Avg Non LGBTQ+ respectively.

Method	Author		CROSSEM-CV		CROSSEM-BERT	
	Rating	Rank	Score	Rank	Score	Rank
LexRank	5.686	4	0.318	5	0.635	5
LSA	6.412	1	0.356	2	0.698	3
BERTSUM	5.945	3	0.339	4	0.720	2
Llama3-70b	5.467	5	0.355	3	0.679	4
GPT-4o-mini	6.226	2	0.388	1	0.748	1

Table 8. Author Rating, CROSSEM score and their corresponding ranks achieved by summarization algorithms for LGBTQ+ survey. Author ratings are closely aligned with the CROSSEM scores.

surveyed on the question: *"Should school students be taught about LGBTQ+ topics, such as gender identity, sexual orientation, and same-sex families, as part of their curriculum?"* Subsequently, 30 human summarizers – 15 from the LGBTQ+ community and 15 from the non-LGBTQ+ community – were tasked with selecting 20 textual units (out of the 200 responses) that they considered most relevant for inclusion in a summary. These selections were treated as reference summaries. In parallel, five different summarization algorithms (as described in Section 3) were used to generate the corresponding algorithmic summaries. Finally, all 200 participants were asked to rate the five algorithmic summaries to assess their level of satisfaction with each summary.

Method	Spearman ρ	Kendall τ
ROUGE	-0.8	-0.6
BERTScore	-0.2	-0.2
MoverScore	-0.2	-0.2
CROSSEM-CV	0.6	0.4
CROSSEM-BERT	0.6	0.4

Table 9. Spearman's ρ and Kendall's τ correlation coefficients computed between different summary evaluation measures on LGBTQ+ survey. ROUGE, BERTScore and MoverScore are negatively correlated with average author ratings; whereas CROSSEM shows positive correlation with the average author ratings across all three surveys.

The results indicate that the reference summaries selected by the summarizers from different communities vary significantly (refer Table 7). The rankings obtained by the algorithmic summaries based on *Rank LGBTQ+* and *Rank Non LGBTQ+* exhibit notable differences. For example, according to ROUGE, LSA emerges as the top-performing summary when only the reference summaries from Non-LGBTQ+ summarizers are considered, whereas Llama achieves the highest rank when evaluations are based on LGBTQ+ summarizers. Similar variations between *Rank LGBTQ+* and *Rank Non LGBTQ+* are evident for BERTScore and MoverScore. These findings further support our earlier hypothesis that reference summaries can differ substantially depending on the opinion group of the summarizers.

Author ratings and the CROSSEM scores computed using two variants – Count Vectorizer and BERT embeddings – are presented in Table 8. The summary produced by GPT-4o-mini is ranked highest by CROSSEM and is also considered the second-best by the authors. To evaluate the alignment between author ratings and different summary evaluation approaches, Spearman's ρ and Kendall's τ correlation coefficients are reported in Table 9. The findings indicate that ROUGE, BERTScore, and MoverScore exhibit negative correlations with author ratings, whereas CROSSEM shows a positive correlation, underscoring its ability to better reflect author satisfaction. Overall, these results suggest that existing summary evaluation metrics are limited in capturing author satisfaction, thereby motivating the need for a new metric such as CROSSEM.

Our experiments across varied participant pools and demographic groups demonstrate that the proposed method effectively capture both reader and author satisfaction. We will further illustrate their applicability on social media datasets in Section 6.

5.4 Towards individual fairness

We demonstrated that the CROSSEM metric can serve as a reliable proxy for gauging author satisfaction with a generated summary. For a given summary $\mathcal{S}$, CROSSEM computes the mean of individual satisfaction scores $sat(a_i, \mathcal{S})$. While this averaging approach provides a convenient way to summarize collective satisfaction, it has a critical limitation: it does not account for disparities in group representation among the authors. For instance, if male authors significantly outnumber female authors, and if the summary disproportionately reflects male favored content, it could still achieve a high CROSSEM score — because the larger group (males) is more satisfied on average. This masks the dissatisfaction of underrepresented groups, such as female authors, leading to biased evaluation outcomes. Consequently, summaries optimized solely for CROSSEM might favor preferences of the majority group and neglect fairness across the author population.

To address this issue and mitigate dominance of any one group's opinion, we propose incorporating individual fairness into the evaluation framework. Our goal is to encourage summaries that ensure uniform satisfaction levels across all authors, regardless of their group affiliations. To operationalize this, we employ a methodology analogous to that presented in Section 4, and augment the CROSSEM metric with the Gini coefficient to capture distributional disparities. Accordingly, the original formulation of CROSSEM (Equation 7) gets modified to a new

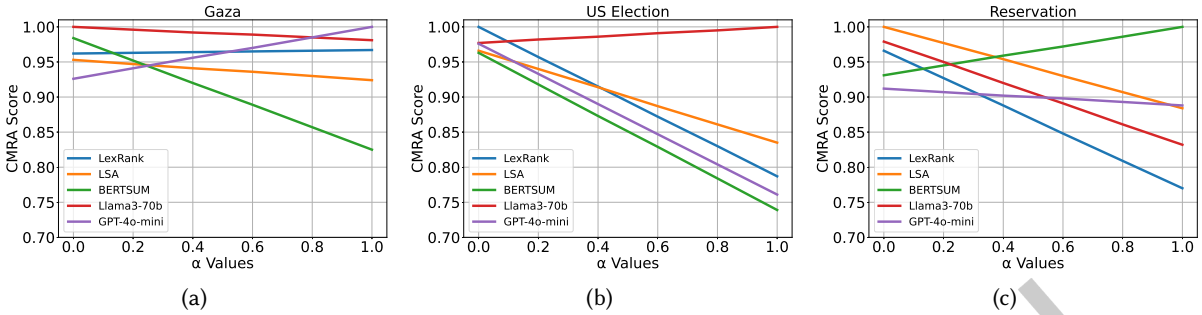


Fig. 6. Plot of CMRA versus α for different surveys (a) Gaza (b) US-Election and (c) Reservation. Change in CMRA scores with respect to α is smooth and gradual; absence of sudden transitions highlights the stability of the CMRA metric.

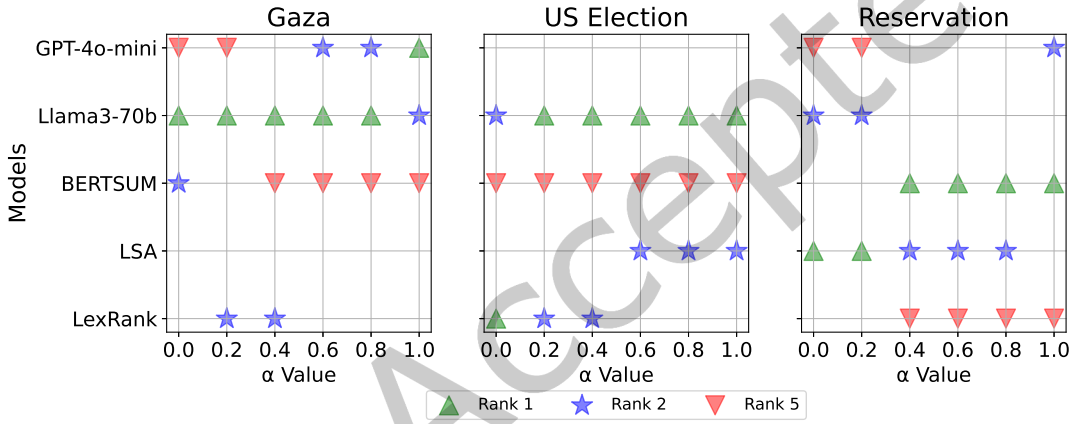


Fig. 7. Plot showing the best, second best and worst performing model for different values of α. Model rankings remain consistent over a wide range of α, showcasing that the CMRA is not sensitive to small changes.

version (Equation 13), which incorporates the inequality.

$$\text{CROSSEM}_g(\mathcal{S}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|sat(a_i,\mathcal{S}) - sat(a_j,\mathcal{S})|}{2\cdot N\cdot\sum_{j=1}^{N} sat(a_j,\mathcal{S})} \tag{13}$$

where N is the number of authors. In our setup, we aim to prioritize algorithmic summaries that achieve the lowest CROSSEM_g value, as this reflects more uniform satisfaction among authors.

5.5 Balancing reader and author satisfaction

As mentioned earlier, the underlying rationale of having human-generated reference summaries in the evaluation setup is to get summaries that would be most helpful to the readers, and hence ROUGE and similar metrics are important in their own right. Our main reservation is in solely using metrics such as ROUGE thereby completely disregarding the author side. Consequently, the idea behind CROSSEM is to capture author satisfaction i.e., how an author perceives a summary, whether it reflects their opinion or not. In this paper, we propose to look at summarization of user-generated data as a two-sided problem where both readers and authors need to be

treated fairly. Thus, for a given algorithmic summary $\mathcal{S}$, we define CMRA (**C**ombined **M**etric for **R**eaders and **A**uthors) score as a linear combination of ROUGE and CROSSEM as shown in equation 14. $f(ROUGE(\mathcal{S}))$ and $f(\texttt{CROSSEM}(\mathcal{S}))$ represents normalized ROUGE and CROSSEM score for a given summary $\mathcal{S}$.

$$\texttt{CMRA}(\mathcal{S}) = \alpha * f(ROUGE(\mathcal{S})) + (1-\alpha) * f(\texttt{CROSSEM}(\mathcal{S})) \tag{14}$$

Max-normalization can be used as the normalizing function. The value of α can be selected based on the desired emphasis between reader satisfaction and author satisfaction. Higher value of α indicates more emphasis on ROUGE (reader satisfaction) and a lower value implies more importance to CROSSEM (author satisfaction). Since these summarization algorithms are often used by some downstream tasks in online platform, the platform designer can set an appropriate value of α depending on the task.

Sensitivity analysis of α: Figure 6 illustrates that CMRA score changes gradually with α across all three datasets, indicating a smooth shift between reader-centric and author-centric evaluation. Since there are no sudden jumps or discontinuities, this establishes the stability of our metric. We also examine the best, second-best, and worst-performing models across different values of α (refer Figure 7). Rankings remain unchanged across all datasets when α lies in the interval [0.6, 0.8], The top-ranked models remain stable over a wide range of α; for Gaza, US-Election, and Reservation datasets, the best-performing model stays consistent over the ranges [0.0, 0.8], [0.2, 1.0] and [0.4, 1.0] respectively. This demonstrates that a platform moderator can check for the extreme value of α and then decide the trade-off between the two objectives instead of checking for each intermediate value of α.

In equation 14, normalized ROUGE and CROSSEM score can even be replaced with normalized ROUGE_g and $\texttt{CROSSEM}_\text{g}$ respectively to account for inequality amongst the readers and authors as shown in equation below.

$$\texttt{CMRA}_\text{g}(\mathcal{S}) = \alpha * f(\text{ROUGE}_\text{g}(\mathcal{S})) + (1-\alpha) * f(\texttt{CROSSEM}_\text{g}(\mathcal{S})) \tag{15}$$

6 Applying CROSSEM in the Wild

So far, we applied CROSSEM only on the controlled survey setup. However, it can easily be used on social media contents like tweets and Facebook posts. To demonstrate the effectiveness of our proposed metrics we specifically use the following three microblog datasets on polarizing topics.

(1) **US Election dataset**: This dataset, provided by [20], contains English tweets posted during the 2016 US Presidential elections. Each tweet is annotated to indicate whether it supports Donald Trump or Hillary Clinton, or remains neutral. For the purpose of analysis, the tweets were consolidated into three opinion categories: (i) Pro-Republican, comprising tweets that either express support for Trump or criticism of Clinton; (ii) Pro-Democratic, encompassing tweets that either express support for Clinton or criticism of Trump; and (iii) Neutral, including tweets that are impartial or critical of both candidates. Following the removal of duplicate tweets, the final dataset consists of 2,120 tweets, of which 1,309 are classified as Pro-Republican, 658 as Pro-Democratic, and 153 as Neutral.

(2) **Claritin dataset**: This dataset contains tweets in English about the effects of the drug Claritin [21]. Each tweet in the dataset is annotated with the gender of the user (male, female, or unknown) who posted it. Tweets associated with users of unknown gender were excluded from the analysis. The final dataset comprises 4,037 tweets, of which 1,532 are authored by male users and 2,505 are authored by female users.

(3) **Me Too dataset**: This dataset [21] contains 10,000 English tweets related to MeToo movement which happened in October 2018 . The dataset comprises tweets authored by male users, female users, and organizational accounts, primarily representing news media agencies. On eliminating duplicate entries and tweets originating from organizational accounts, the dataset was reduced to 488 tweets, comprising 213 authored by male users and 275 by female users.

Method	ROUGE		$ROUGE_g$		CROSSEM		$CROSSEM_g$	
	Score	Rank	Score	Rank	Score	Rank	Score	Rank
US Election								
ClusterRank	0.497	4	0.017	7	0.093	2	0.233	11
DSDR	0.510	2	0.009	1	0.080	9	0.210	4
LexRank	0.496	5	0.025	11	0.085	4	0.223	8
SummBasic	0.498	3	0.018	8	0.083	6	0.233	11
LSA	0.518	1	0.010	2	0.083	6	0.200	1
LUHN	0.492	7	0.032	12	0.100	1	0.231	9
SummaRunner	0.494	6	0.016	6	0.078	11	0.211	5
BERTSUM	0.431	11	0.011	3	0.084	5	0.215	6
Llama3-70b	0.446	8	0.021	10	0.081	8	0.208	3
Mixtral-8x7b	0.424	12	0.015	5	0.078	11	0.218	7
Gemini-1.0-pro	0.434	10	0.019	9	0.080	9	0.207	2
GPT-4o-mini	0.438	9	0.014	4	0.088	3	0.232	10
Claritin								
ClusterRank	0.361	10	0.010	3	0.128	12	0.127	4
DSDR	0.438	5	0.014	4	0.166	4	0.148	8
LexRank	0.346	12	0.021	9	0.243	1	0.173	11
SummBasic	0.383	9	0.017	6	0.174	2	0.203	12
LSA	0.393	8	0.014	4	0.133	11	0.109	1
LUHN	0.358	11	0.009	2	0.134	10	0.129	5
SummaRunner	0.404	7	0.007	1	0.148	9	0.126	3
BERTSUM	0.444	4	0.020	7	0.153	8	0.135	6
Llama3-70b	0.490	2	0.020	7	0.165	6	0.135	6
Mixtral-8x7b	0.473	3	0.024	11	0.173	3	0.152	9
Gemini-1.0-pro	0.424	6	0.023	10	0.166	4	0.167	10
GPT-4o-mini	0.507	1	0.028	12	0.159	7	0.125	2
Me Too								
ClusterRank	0.441	4	0.006	5	0.159	5	0.161	11
DSDR	0.436	5	0.015	9	0.157	6	0.153	8
LexRank	0.415	10	0.014	8	0.160	4	0.151	6
SummBasic	0.414	11	0.016	10	0.162	3	0.152	7
LSA	0.395	12	0.004	3	0.152	9	0.135	1
LUHN	0.433	8	0.010	6	0.169	2	0.168	12
SummaRunner	0.496	1	0.003	2	0.172	1	0.150	5
BERTSUM	0.463	3	0.005	4	0.150	10	0.159	10
Llama3-70b	0.436	5	0.029	11	0.145	12	0.145	2
Mixtral-8x7b	0.436	5	0.001	1	0.146	11	0.145	2
Gemini-1.0-pro	0.473	2	0.011	7	0.154	7	0.149	4
GPT-4o-mini	0.431	9	0.039	12	0.153	8	0.158	9

Table 10. ROUGE, CROSSEM, $ROUGE_g$ and $CROSSEM_g$ score for various social media datasets. For ROUGE and CROSSEM, a higher score is desirable, whereas for inequality based $ROUGE_g$ and $CROSSEM_g$ a lower score is preferred. For inequality based methods, a lower score denotes equality i.e., all the values are nearly the same, while a score of 1 reflects maximal inequality among values.

All three datasets contain tweets that express various opinions of different groups of authors (male and female, or authors of different political leanings). While summarizing such user generated data, we assume that the authors will be satisfied if their opinions are reflected in the summary. Table 10 shows results of ROUGE, CROSSEM, ROUGE_g and CROSSEM_g for social media datasets - US Election, Claritin and Me Too.

In US Election dataset, LSA delivers highest reader satisfaction (ROUGE score), whereas LUHN ensures utmost author satisfaction (CROSSEM score). Table 10 demonstrates that an algorithm, rated as the top performer in ROUGE scores, may not necessarily be the best choice if author satisfaction is the primary factor to consider. For CROSSEM_g, DSDR is the top rated algorithm, indicating that the CROSSEM score achieved by comparing DSDR summary with various textual units written by authors, are almost identical. Thus, we can conclude that an algorithm that exhibit impressive ranking in terms of ROUGE score, can still fall short of being the optimal choice for author satisfaction.

7 Conclusion

In this paper, we demonstrated that the traditional text summarization evaluation setup – which hinges on comparing candidate summaries with reference summaries created by human annotators – may not be suitable for evaluating summaries generated from user generated text on polarizing socio-political topics. This is due to two problems – i). Inherent bias of the summarizers is likely to result in very different reference summaries for the same input text, which might lead to the preference of a biased algorithmic summary based on metrics like ROUGE, BERTScore and MoverScore. ii). Conventional metrics for text summarization do not consider the satisfaction of the author, who are key stakeholders in the summarization process as it is their viewpoints that are being summarized. Hence, for evaluating summaries generated from user-generated text, we look at summarization from both readers' and authors' perspective. To address the first problem – ensuring fairness to readers by selecting an equitable algorithmic summary – we propose ROUGE_g, an inequality-based variation of ROUGE, which reflects the uniformity in the distribution of ROUGE score. For second problem, we introduce CROSSEM, a novel evaluation measure that explicitly accounts for author satisfaction. To ensure individual fairness among authors, we rely on inequality-based index measure – CROSSEM_g. We can extend the proposed framework by using CMRA, which is a linear combination of ROUGE and CROSSEM where the weights can be decided according to one's preference. One advantage of our evaluation method CROSSEM is that it is not dependent on reference summaries, and hence can work even in situations where human written summaries are not available. Aim of this work is to propose a new evaluation approach for evaluating summaries of user generated textual units which considers fairness towards both readers and authors.

Note that CROSSEM does not require authors to explicitly provide their preferred subset of textual units in real-world deployments. The author's approved set consist of the textual units written by the author. Semantic similarity is leveraged to find the similarity between the generated summary and the original textual units. Higher semantic alignment indicates that the summary better reflects the content and intent of authors. At the same time, we acknowledge that semantic similarity is an imperfect proxy for true author satisfaction. It primarily captures content overlap, but may overlook other dimensions such as tone, contextual framing, emphasis, or potential misrepresentation of an author's viewpoint. Therefore, CROSSEM should be viewed as a foundational step toward modeling author satisfaction rather than a complete solution. Future extensions could address these limitations by incorporating richer representations of similarity. For instance, a Small Language Model (SLM) can be trained to capture multiple facets of alignment – including semantic fidelity, sentiment consistency, and contextual appropriateness – and its outputs can be integrated into CROSSEM to provide a more comprehensive and nuanced estimation of author satisfaction.

The metrics proposed in Section 5 measure mean satisfaction of all authors. We propose this intuitive approach as a proof of concept in this paper. The said metrics can be tweaked in multiple different ways to account for

multiple other nuances, optimizing for which, even fairness in text summarization can also be achieved. For example, replacing the mean function with a max-min function, one can evaluate a summary that maximizes the minimum satisfaction of individual authors. The proposed metric can be employed to monitor distributional shifts. For example, during real-time summary generation, author satisfaction levels can be tracked, and if they fall below a minimum threshold, the summarization process can be reinitiated. Further, we acknowledge that the end summaries should not contain any misinformation or hateful content towards any group or individual of the society. An extensive line of research has developed in recent years, focusing on content moderation to mitigate these concerns [1, 38]. Once user-generated texts have passed moderation filters, summarization algorithms can then be employed to generate summaries.

Future directions: We believe that our proposed evaluation approach will initiate a new discussion on how to properly evaluate the summarization algorithms especially for user generated textual units. In future, we plan to apply the proposed approach to abstractive summaries and investigate more nuanced and efficient methods of capturing authors' preferences. We also plan to focus on generating summaries with a primary objective as maximizing the author satisfaction.

References

[1] Sayantan Adak, Souvic Chakraborty, Paramtia Das, Mithun Das, Abhisek Dash, Rima Hazra, Binny Mathew, Punyajoy Saha, Soumya Sarkar, and Animesh Mukherjee. 2021. Mining the online infosphere: A survey. *arXiv preprint arXiv:2101.00454* (2021).

[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut. 2017. Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications* (2017).

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica* 23 (2016).

[4] Sina Bagheri Nezhad, Sayan Bandyapadhyay, and Ameeta Agrawal. 2025. Fair Summarization: Bridging Quality and Diversity in Extractive Summaries. In *C3NLP*.

[5] Irene Benedetto, Moreno La Quatra, Luca Cagliero, Luca Vassio, and Martino Trevisan. 2024. TASP: Topic-based abstractive summarization of Facebook text posts. *Expert Systems with Applications* (2024).

[6] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. 2019. Fair algorithms for clustering. *NeurIPS* (2019).

[7] Arpita Biswas, Gourab K. Patro, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2021. Toward Fair Recommendation in Two-Sided Platforms. *ACM Trans. Web* (2021).

[8] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. 2017. Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics* 3, 1 (2017).

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[10] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *ACM FAccT*.

[11] Luis Adrián Cabrera-Diego and Juan-Manuel Torres-Moreno. 2018. SummTriver: A new trivergent model to evaluate summaries automatically without human references. *Data & Knowledge Engineering* (2018).

[12] Luis Adrián Cabrera-Diego, Juan-Manuel Torres-Moreno, and Barthélémy Durette. 2016. Evaluating Multiple Summaries Without Human Models: A First Experiment with a Trivergent Model. In *Natural Language Processing and Information Systems*.

[13] Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth Vishnoi. 2018. Fair and diverse DPP-based data summarization. In *ICML*.

[14] Abhijnan Chakraborty, Johnnatan Messias, Fabricio Benevenuto, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. 2017. Who makes trends? understanding demographic biases in crowdsourced recommendations. In *AAAI ICWSM*.

[15] Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. 2019. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *ACM FAccT*.

[16] Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. BooookScore: A systematic exploration of book-length summarization in the era of LLMs. In *ICLR*.

[17] Garima Chhikara, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2023. Fairness for Both Readers and Authors: Evaluating Summaries of User Generated Content. In *SIGIR*.

[18] Garima Chhikara, Anurag Sharma, V. Gurucharan, Kripabandhu Ghosh, and Abhijnan Chakraborty. 2025. LaMSUM: Amplifying Voices Against Harassment through LLM Guided Extractive Summarization of User Incident Reports. arXiv:2406.15809

[19] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair Clustering Through Fairlets. In *NIPS*.

[20] K. Darwish, W. Magdy, and Zanouda T. 2017. Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election. In *SocInfo*.

[21] 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. *ACM CSCW* (2019).

[22] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. 2021. Two-sided fairness in rankings via Lorenz dominance. In *NeurIPS*.

[23] Robert L. Donaway, Kevin W. Drummey, and Laura A. Mather. 2000. A Comparison of Rankings Produced by Summarization Evaluation Measures. In *NAACL-ANLP*.

[24] Yue Dong. 2018. A Survey on Neural Network-Based Summarization Methods. *arXiv preprint arXiv:1804.04589* (2018).

[25] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications* (2021).

[26] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *Journal of Artificial Intelligence Research* (2004).

[27] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the ACL* (2021).

[28] Ruoyuan Gao and Chirag Shah. 2020. Counteracting Bias and Increasing Fairness in Search and Recommender Systems. In *ACM RecSys*.

[29] Ruoyuan Gao and Chirag Shah. 2021. Addressing Bias and Fairness in Search Systems. In *ACM SIGIR*.

[30] Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani-Tür. 2009. Clusterrank: a graph based method for meeting summarization. In *Proc. Interspeech*.

[31] Google Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint* (2023).

[32] Dan Gillick and Yang Liu. 2010. Non-Expert Evaluation of Summarization Systems is Risky. In *NAACL HLT*.

[33] Corrado Gini. 1912. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.* Tipogr. di P. Cuppini.

[34] Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *ACM SIGIR*.

[35] Andrew Griffin. 2015. Twitter and Google team up, so tweets now go straight into Google search results. https://www.independent.co.uk/tech/twitter-and-google-team-up-so-tweets-now-go-straight-into-google-search-results-10262417.html.

[36] Vishal Gupta and Gurpreet Singh Lehal. 2010. A Survey of Text Summarization Extractive Techniques. *IEEE Journal of Emerging Tech. in Web Intelligence* (2010).

[37] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *ACM KDD*.

[38] Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Commun. ACM* (2022).

[39] Zhanying He, Chun Chen, Jiajun Bu, Can Wang, Lijun Zhang, Deng Cai, and Xiaofei He. 2012. Document Summarization Based on Data Reconstruction. In *AAAI*.

[40] Run Huang, Anna Katherine Zhao, Zeinabsadat Saghi, Sadra Sabouri, and Souti Chattopadhyay. 2025. Beyond the Page: Enriching Academic Paper Reading with Social Media Discussions. In *UIST*.

[41] Andrew Hutchinson. 2015. The First Examples of Real-Time Tweets in Google Search Have Arrived. https://www.linkedin.com/pulse/first-examples-real-time-tweets-google-search-have-andrew-hutchinson/.

[42] David I. Inouye and Jugal K. Kalita. 2011. Comparing Twitter Summarization Algorithms for Multiple Post Summaries.. In *IEEE SocialCom*.

[43] Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*.

[44] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. arXiv:2401.04088

[45] Wenjun Jiang, Jing Chen, Xiaofei Ding, Jie Wu, Jiawei He, and Guojun Wang. 2021. Review Summary Generation in Online Systems: Frameworks for Supervised and Unsupervised Scenarios. *ACM Trans. Web* (2021).

[46] Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. arXiv:2403.02901

[47] Shawn M. Jones, Martin Klein, Michele C. Weigle, and Michael L. Nelson. 2023. Summarizing Web Archive Corpora via Social Media Storytelling by Automatically Selecting and Visualizing Exemplars. *ACM Trans. Web* (2023).
[48] Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A Comprehensive Survey of Sentence Representations: From the BERT Epoch to the ChatGPT Era and Beyond. In *EACL*.
[49] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* (1938).
[50] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From Word Embeddings To Document Distances. In *ICML*.
[51] Martin Lackner, Peter Regner, and Benjamin Krenn. 2023. abcvoting: A Python package for approval-based multi-winner voting rules. *Journal of Open Source Software* (2023).
[52] Haoyuan Li, Yusen Zhang, Rui Zhang, and Snigdha Chaturvedi. 2025. Coverage-based Fairness in Multi-document Summarization. In *NAACL*.
[53] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out, ACL*.
[54] Yang Liu and Mirella Lapata. 2019. Text Summarization with Pretrained Encoders. In *EMNLP*.
[55] Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On Learning to Summarize with Large Language Models as References. In *NAACL*.
[56] Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. In *COLING*.
[57] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* (1958).
[58] S. Mackie, R. McCreadie, C. Macdonald, and I. Ounis. 2014. Comparing Algorithms for Microblog Summarisation. In *CLEF*.
[59] Samaneh Mohtadi, Kevin Roitero, Stefano Mizzaro, and Gianluca Demartini. 2026. The Effect of Document Summarization on LLM-Based Relevance Judgments. In *Advances in Information Retrieval*.
[60] Aadi Swadipto Mondal, Rakesh Bal, Sayan Sinha, and Gourab K Patro. 2021. Two-Sided Fairness in Non-Personalised Recommendations (Student Abstract). In *AAAI*.
[61] Mohammed Elsaid Moussa, Ensaf Hussein Mohamed, and Mohamed Hassan Haggag. 2018. A survey on opinion summarization techniques for social media. *Future Computing and Informatics Journal* (2018).
[62] Rajdeep Mukherjee, Hari Chandana Peruri, Uppada Vishnu, Pawan Goyal, Sourangshu Bhattacharya, and Niloy Ganguly. 2020. Read what you need: Controllable aspect-based opinion summarization of tourist reviews. In *ACM SIGIR*.
[63] Rajdeep Mukherjee, Uppada Vishnu, Hari Chandana Peruri, Sourangshu Bhattacharya, Koustav Rudra, Pawan Goyal, and Niloy Ganguly. 2022. MTLTS: A Multi-Task Framework To Obtain Trustworthy Summaries From Crisis-Related Microblogs. In *ACM WSDM*.
[64] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI*.
[65] Ani Nenkova and Lucy Vanderwende. 2005. *The impact of frequency on summarization.* Technical Report. Microsoft Research.
[66] Library of Congress. 2017. Update on the Twitter Archive at the Library of Congress. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_whitepaper.pdf.
[67] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.
[68] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NIPS*.
[69] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *WWW*.
[70] Gourab K. Patro, Abhijnan Chakraborty, Niloy Ganguly, and Krishna Gummadi. 2020. Incremental Fairness in Two-Sided Market Platforms: On Smoothly Updating Recommendations. In *AAAI*.
[71] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *ACM KDD*.
[72] Prolific. 2024. Can I screen participants within my study? https://researcher-help.prolific.com/en/article/4ae222.
[73] Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (Almost) Dead. arXiv:2309.09558
[74] Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Çelebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation Challenges in Large-Scale Document Summarization. In *ACL*.
[75] Koustav Rudra, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh. 2018. Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters. *ACM Trans. Web* (2018).
[76] Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, and Eric SanJuan. 2010. Multilingual Summarization Evaluation without Human Models. In *COLING*.
[77] David Sayce. 2022. The Number of tweets per day in 2022. https://www.dsayce.com/social-media/tweets-day/.
[78] Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2018. Summarization Evaluation in the Absence of Human Model Summaries Using the Compositionality of Word Embeddings. In *COLING*.
[79] Anurag Shandilya, Abhisek Dash, Abhijnan Chakraborty, Kripabandhu Ghosh, and Saptarshi Ghosh. 2020. Fairness for Whom? Understanding the Reader's Perception of Fairness in Text Summarization. In *FILA workshop, IEEE Big Data*.

[80] Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of Extractive Text Summarization. In *Companion Proceedings of WWW*.
[81] B. Sharifi, M. Hutton, and J. K. Kalita. 2010. Experiments in Microblog Summarization. In *IEEE Conference on Social Computing*.
[82] C. Spearman. 1987. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* (1987).
[83] Anna Squicciarini, Sarah Rajtmajer, Yang Gao, Justin Semonsen, Andrew Belmonte, and Pratik Agarwal. 2022. An Extended Ultimatum Game for Multi-Party Access Control in Social Networks. *ACM Trans. Web* (2022).
[84] Josef Steinberger and Karel Jezek. 2009. Evaluation Measures for Text Summarization. *Computing and Informatics* (2009).
[85] Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A Simple Framework for Opinion Summarization. In *ACL*.
[86] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin Rousseau, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *medRxiv* (2023).
[87] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.
[88] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2022. A Multi-Objective Optimization Framework for Multi-Stakeholder Fairness-Aware Recommendation. *ACM Transactions on Information Systems* (2022).
[89] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A Two-Sided Fairness-Aware Recommendation Model for Both Customers and Providers. In *ACM SIGIR*.
[90] Wei Xu, Ralph Grishman, Adam Meyers, and Alan Ritter. 2013. A Preliminary Study of Tweet Summarization using Information Extraction. In *Language in Social Media (LASM)*.
[91] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *ACM CIKM*.
[92] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *Comput. Surveys* (2022).
[93] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.
[94] Yanyue Zhang, Pengfei Li, Yilong Lai, Yulan He, and Deyu Zhou. 2025. LASS: A Novel and Economical Data Augmentation Framework Based on Language Models for Debiasing Opinion Summarization. In *COLING*.
[95] Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, Kathleen McKeown, and Rui Zhang. 2023. Fair Abstractive Summarization of Diverse Perspectives. arXiv:2311.07884
[96] Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *EMNLP-IJCNLP*.

Received 26 April 2025; revised 16 April 2026; accepted 8 July 2026