Fairness for both Readers and Authors: Evaluating Summaries of User Generated Content

Garima Chhikara
Indian Institute of Technology Delhi
Delhi Technological University
New 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 Delhi
New Delhi, India

ABSTRACT

Summarization of textual content has many applications, ranging from summarizing long documents to recent efforts towards summarizing user generated text (e.g., tweets, Facebook or Reddit posts). Traditionally, the focus of summarization has been to generate summaries which can best satisfy the readers. In this work, we look at summarization of user-generated content as a two-sided problem where satisfaction of both readers and authors is crucial. Through three surveys, we show that for user-generated content, traditional evaluation approach of measuring similarity between reference summaries and algorithmic summaries cannot capture author satisfaction. We propose an author satisfaction-based evaluation metric `CROSSEM` which, we show empirically, can potentially complement the current evaluation paradigm. We further propose the idea of inequality in satisfaction, to account for individual fairness amongst readers and authors. To our knowledge, this is the first attempt towards developing a fair summary evaluation framework for user generated content, and is likely to spawn lot of future research in this space.

CCS CONCEPTS

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

KEYWORDS

Fair Summarization, Summary Evaluation, Author Satisfaction

ACM Reference Format:
Garima Chhikara, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2023. Fairness for both Readers and Authors: Evaluating Summaries of User Generated Content. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591986

1 INTRODUCTION

Recent explosion in the amount of content available on internet has necessitated algorithms which can automatically deliver accurate summaries [12, 22]. A large body of prior research has focused on summarizing (single) long documents such as news articles [1, 11]. However, in recent years, summarization has been increasingly applied on different types of *user generated content* (e.g., tweets, Reddit posts) [8, 24, 31, 35, 45, 49], where the task is to summarize short, independent textual posts written by many authors [33]. There is a fundamental difference between these two summarization tasks. In the latter case, the input text can contain a much wider variety of opinions expressed by thousands of different *authors*, compared to summarizing one or a small set of long documents. Especially, if the posts are on a polarizing topic such as politics, gender-related issues, religion, etc., then the posts can contain mutually conflicting opinions, and it is natural to ask whether the different opinions are *fairly reflected* in the summary [8].

The prevalent evaluation setup for text summarization involves comparing the *algorithmic summary* (generated by a summarization algorithm) with *reference summaries* (also known as *gold-standard summaries*) written by human beings (whom we refer to as *summarizers*), using standard evaluation measures like ROUGE [27]. These reference summaries act as proxy for what the eventual *readers* would want from a summary, and thus, evaluation metrics like ROUGE essentially attempt to capture reader satisfaction. An algorithmic summary which is highly correlated with the human-written summary is considered to be the best for the readers.

However, such reference summary based evaluation, if applied in the context of summarizing user generated content, completely ignores the authors whose opinions are being summarized. Summarization provides visibility to the posts 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 [21]; Library of Congress only stores a selection of tweets as part of its Twitter Archive while emphasizing the importance of giving voice to the common people [38]. Thus, *summarization of user generated content is essentially a two-sided problem*, and to be fair to both sides, alongside reader satisfaction, author satisfaction also needs to be considered. To this end, in this work, we propose a new evaluation metric `CROSSEM` (CROwd Satisfaction-based Summary Evaluation Metric) which attempts to capture author satisfaction, and can nicely complement the existing evaluation paradigm.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591986

Topic	Country	Opinion Group
Is the current economic condition of United Kingdom caused by **Brexit** ?	UK	Remain Leave
Should Covid-19 **vaccine** be made mandatory for everyone?	USA	Pro-vaccine Anti-vaccine
Do Indian IT companies prefer hiring female students over male students during on-campus **recruitment?**	India	Male Female

Table 1: We conducted three surveys on three controversial topics with participants from the UK, USA and India.

Contributions: We make the following contributions in this paper:

- We design a novel survey (§2) where we first ask authors to write their opinions on three topics, and get them summarized by human summarizers as well as by 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. We observe that reference summary based evaluation is inadequate to capture author satisfaction.
- We propose an author satisfaction based evaluation metric CROSSEM which we found to be highly correlated with author provided scores. Furthermore, 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 further look into the *individual fairness* among readers and authors, and propose extensions of ROUGE and CROSSEM which consider inequality in satisfaction.
- To our knowledge, ours is the first work to consider summarization of user generated content as a two-sided problem and the fairness therein. While we focus on evaluating summaries, it should spawn future works looking to produce two-sided fair summaries.

Related Works: Several recent works have focused on two-sided notion of fairness for producers and consumers in recommendation [2, 3, 7, 32, 39, 40, 47, 48] and search systems [9, 15, 16]. In this work, we introduce the idea of two-sided fairness during summarization of user-generated content. Prior works have explored summary evaluation in absence of reference summaries, albeit for long documents [4, 5, 10, 29, 41–43]. While our proposal CROSSEM is driven by similar idea, the motivation is from the fairness perspective. Moreover, a metric focusing on individual satisfaction is a novel contribution in itself. Finally, while some approaches have attempted to generate fair summaries [6, 8, 34], concerns have been raised regarding the efficacy of the evaluation approaches [44]. The current work is an attempt to bridge such gaps in the literature.

2 CAN REFERENCE SUMMARY BASED EVALUATION SATISFY AUTHORS?

In this section, through surveys conducted with participants from three different countries, we investigate whether reference summary based evaluation is sufficient to capture author satisfaction.

2.1 Survey setup

We chose three debatable topics to get responses from participants having different opinions on them. Table 1 presents the details about the surveys conducted with participants from UK, USA and India. We used Prolific (www.prolific.co) for surveys on Brexit and Vaccine. For each survey, we selected equal number of participants belonging to two opinion groups, where the opinions were extracted based on their answers to a pre-screening question. For example, for Brexit, 18 participants voted in favor of UK remaining in the European Union (EU) and the other 18 voted in favour of UK leaving EU. For Vaccine survey, participants had either positive or negative attitude towards the covid vaccines. For Recruitment, we could not use Prolific as the number of available respondents from India were not adequate. Hence, we conducted the survey with final year undergraduate students from an Indian University, who are seeking jobs through on-campus placement interviews, and thus the topic is very relevant to them.[1] To get the perceptions of different gender groups, 18 male and 18 female students were chosen for 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 1.

Phase I: We ask 30 participants to write 6-8 sentences on a particular topic. We call these participants *authors*. From Phase I, we receive a set of textual units written by authors for each topic.

Phase II: We then apply seven extractive summarization algorithms – ClusterRank [17], DSDR [23], LexRank [13], LSA [20], LUHN [30], SumBasic [37], SummaRuNNer-RNN [36] – on the textual units to obtain several algorithmic summaries. The prevailing method to measure the goodness of algorithmic summaries is to compare them with human written reference summaries. Thus, we ask 6 additional participants (3 each from every opinion group) to create summaries by selecting 10-12 most relevant textual units (this process is known as *extractive summarization*). We call these participants as *summarizers*. Using the reference summaries, we use three different evaluation metrics ROUGE [27], BertScore [50] and MoverScore [51] to rank different summarization algorithms. Table 2 shows ROUGE and BertScore between different algorithmic and reference summaries for Vaccine survey. Results for MoverScore and other surveys are omitted due to lack of space. Note that the same approach can be easily extended to any evaluation metric dependent on reference summaries.

2.2 Going back to the authors

The underlying motivation behind including the reference summaries in the evaluation process is to assess how well an algorithm generated summary would be useful to the readers. However, in summarization of user generated content, alongside the readers, authors are also important stakeholders. Considering them is especially relevant for polarizing topics, since authors of different viewpoints would like their opinions to be reflected in the summary. Thus, rather than relying only on human summarizers with unknown implicit biases, it might be a good idea to go back to the authors to see how they perceive different algorithmic summaries. Thus, in **Phase III** of the surveys, we showed 7 algorithmic

[1] The survey was conducted after taking appropriate institute permissions. All survey participants were informed in advance of the motivation and purpose of the study, and they voluntarily agreed to participate in the survey.

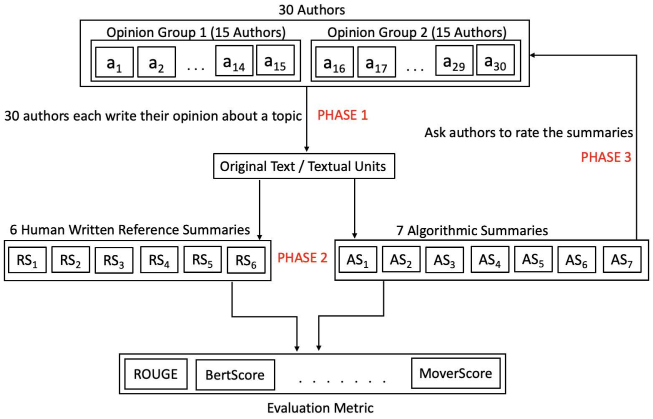

30 Authors
Opinion Group 1 (15 Authors)
Opinion Group 2 (15 Authors)
a1
a2
...
a14
a15
a16
a17
...
a29
a30
30 authors each write their opinion about a topic
PHASE 1
Ask authors to rate the summaries
PHASE 3
Original Text / Textual Units
6 Human Written Reference Summaries
7 Algorithmic Summaries
RS1
RS2
RS3
RS4
RS5
RS6
PHASE 2
AS1
AS2
AS3
AS4
AS5
AS6
AS7
ROUGE
BertScore
. . . .
MoverScore
Evaluation Metric

Figure 1: Diagram representing the flow of the survey. Each survey is conducted in three phases involving authors and summarizers. Each phase in explained in detail in §2.

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.

In Table 2, the *Author* column shows the average rating provided by the authors to various algorithmic summaries and resultant ranking of the algorithms. For `Vaccine` survey, it can be observed that *Author* rankings are different from the rankings obtained through ROUGE and BertScore. For example, SummaRNN achieves the highest ROUGE and BertScore; however, the authors seem to be most satisfied with ClusterRank. We observed similar patterns for other topics as well.

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. However, it is not practical 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 next.

3 AUTHOR SATISFACTION BASED METRIC FOR SUMMARY EVALUATION

In this section, we propose a metrics named CROSSEM (**CRO**wd **S**atisfaction-based **S**ummary **E**valuation **M**etric) which can automatically capture the satisfaction of authors.

3.1 Introducing CROSSEM

The underlying idea behind CROSSEM is that a good summary 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. 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 they 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.[2] With this underlying rationale, we now formally define the author satisfaction-based summary evaluation.

Method	ROUGE		BertScore		Author	
	Rating	Rank	Rating	Rank	Rating	Rank
`Vaccine`						
ClusterRank	0.292	7	0.799	6	**6.188**	1
DSDR	0.307	6	0.791	7	6.186	2
LexRank	0.403	2	0.82	2	5.686	4
SummBasic	0.334	5	0.803	4	4.625	7
LSA	0.345	3	0.806	3	5.188	6
LUHN	0.335	4	0.803	4	5.688	3
SummaRNN	**0.44**	1	**0.828**	1	5.438	5

Table 2: Ranking of different algorithms based on ROUGE, BertScore and Author Scores for `Vaccine` survey. Best scores are highlighted in bold.

Definition 1 (Author Satisfaction-Based Summary Evaluation). *Let $\mathcal{T} = \{t_1, t_2, \ldots t_N\}$ be a set (universe) of textual units, written by a set of authors $\mathcal{A} = \{a_1, a_2, \ldots a_N\}$, where $t_i \in \mathcal{T}$ has been written by $a_i \in \mathcal{A}$. A textual unit can be one post or one sentence. 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. Let $|V_i| = l_i$. 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 $\mathcal{S}$.*

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. Our proposed metric CROSSEM quantifies such goodness of a summary. Different versions of the metric can be defined, by defining the satisfaction measure $\mathrm{sat}(a_i, \mathcal{S})$ in various ways.

Definition 2. CROSSEM *(semantic match-based satisfaction): Here we compute the satisfaction of author a_i based on the semantic similarity between V_i and $\mathcal{S}$*

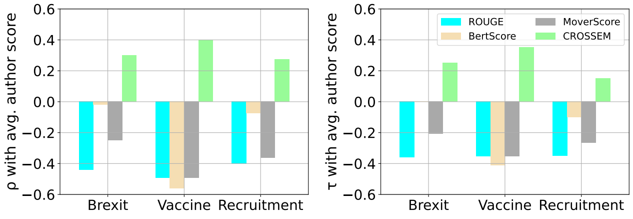
$$sat_s(a_i, \mathcal{S}) = sim(V_i, \mathcal{S}) \tag{1}$$

where $sim(V_i, \mathcal{S}) \in [0, 1]$ is a measure of the semantic similarity between two sets of textual units. CROSSEM *measures the mean (average) satisfaction of all authors for a given summary.*

$$\texttt{CROSSEM}(\mathcal{S}) = \frac{1}{N} \sum_{i=1}^{N} sat_s(a_i, \mathcal{S}) \tag{2}$$

The semantic similarity $sim(V_i, \mathcal{S})$ can be measured in a wide variety of ways. There is a long line of research on measuring the semantic similarity between two pieces of text, ranging from TF-IDF similarity, to using word embeddings and neural models [4, 5, 10, 28, 29, 41–43]. We use cosine similarity for our expermiments.

[2] We understand that, while summarizing crowdsourced text, it is impractical to assume that each author will indicate an approved set of textual units. One easy alternative is to assume that each author only approves the textual units they have written.



(a) Spearman Correlation ρ **(b) Kendall Correlation τ**

Figure 2: (a) Spearman's ρ and (b) Kendall's τ correlation coefficients computed between different summary evaluation measures on Brexit, Vaccine and Recruitment surveys.

Two interesting points can be noted about CROSSEM:
(1) It can work in the absence of any reference summaries, thus can be continuously observed in a real-world summarization setting.
(2) It can be used to evaluate both *extractive and abstractive summaries* by suitably defining the similarity function. In the present work, however, we are only considering extractive summarization. Note that the values of CROSSEM are defined in the range $[0.0, 1.0]$.

3.2 Can CROSSEM satisfy authors in reality?

We check the performance of CROSSEM on the data obtained from three surveys, and compare with ROUGE, BertScore and MoverScore. To quantify the closeness of various metrics to author score, we use Spearman's [46] and Kendall's [26] rank based correlation coefficients. Figures 2(a) and 2(b) show the rank correlation of average author scores with ROUGE, BertScore, MoverScore and CROSSEM. It can be seen that all three reference summary based metrics ROUGE, BertScore and MoverScore are negatively correlated with author scores, while CROSSEM has positive correlation with author scores across all three surveys, demonstrating that it can act as a good proxy of author satisfaction.

3.3 Combining 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. Whereas, the idea behind CROSSEM is to capture author satisfaction i.e., how an author perceive a summary, whether it reflects their opinion or not. 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 3. $f(ROUGE(\mathcal{S}))$ and $f(\text{CROSSEM}(\mathcal{S}))$ provides normalized ROUGE and CROSSEM score respectively for given summary $\mathcal{S}$.

$$\text{CMRA}(\mathcal{S}) = \alpha * f(ROUGE(\mathcal{S})) + (1-\alpha) * f(\text{CROSSEM}(\mathcal{S})) \quad (3)$$

The value of α can be chosen depending on whether we want to put more emphasis on reader satisfaction or author satisfaction, the platform designer can set an appropriate value of α depending on the task. Table 3 shows CMRA score obtained through a linear combination of ROUGE and CROSSEM with $\alpha = 0.5$. We can observe that LUHN is the best algorithm when only author satisfaction (CROSSEM) is considered; whereas LexRank serves as the best algorithm when both readers' and authors' satisfactions are important.

Method	CROSSEM	CMRA $\alpha = 0.5$	ROUGE$_g$	CROSSEM$_g$
Vaccine				
ClusterRank	0.183	0.872	0.161	0.198
DSDR	0.148	0.807	0.174	0.219
LexRank	0.167	**0.944**	**0.06**	**0.135**
SummBasic	0.15	0.883	0.11	0.227
LSA	0.132	0.848	0.089	0.138
LUHN	**0.188**	0.913	0.13	0.191
SummaRNN	0.163	0.931	0.098	0.145

Table 3: CROSSEM, CMRA with $\alpha = 0.5$, ROUGE$_g$ and CROSSEM$_g$ for Vaccine survey. Best scores are highlighted in bold.

3.4 Towards individual fairness

CROSSEM metric, for a given summary $\mathcal{S}$, takes mean over $sat_s(a_i, \mathcal{S})$, it might lead to over-representation of one group. To prevent dominance of one opinion, we propose to *ensure individual fairness among authors*, where a summary would attempt to ensure equal satisfaction of all authors. We use the gini index [19], originally proposed to compute income inequality, alongside CROSSEM to capture inequality in author satisfaction.

$$\text{CROSSEM}_g(\mathcal{S}) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} |sat_s(a_i,\mathcal{S}) - sat_s(a_j,\mathcal{S})|}{2 \cdot N \cdot \sum_{j=1}^{N} sat_s(a_j,\mathcal{S})} \quad (4)$$

Similar to authors, we can also think about individual fairness for readers. Traditionally, an algorithmic summary $\mathcal{S}$ is compared with multiple reference summaries R_i and arithmetic mean over these similarity scores is reported as the final ROUGE score. Multiple prior works have revealed that evaluation based on human written reference summaries is dependent on the choice of summarizers, due to their (often implicit) biases [14, 18, 25]. Similar to CROSSEM$_g$, we can have ROUGE with gini index, as defined below.

$$\text{ROUGE}_g(\mathcal{S}) = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} |ROUGE(R_i,\mathcal{S}) - ROUGE(R_j,\mathcal{S})|}{2 \cdot K \cdot \sum_{j=1}^{K} ROUGE(R_j,\mathcal{S})} \quad (5)$$

Where K is the number of summarizers and $ROUGE(R_i, \mathcal{S})$ is ROUGE score between reference summary R_i and algorithmic summary $\mathcal{S}$.

In Table 3, it can be observed when individual fairness among reader (ROUGE$_g$) and author (CROSSEM$_g$) is considered, LexRank proves to be the best algorithm; CMRA with $\alpha = 0.5$ also reports LexRank as the top rated algorithm. The results from Table 3 build up the credibility of our combined metric CMRA. In equation 3, normalized ROUGE$_g$ and CROSSEM$_g$ can be used instead of normalized ROUGE and CROSSEM to account for individual fairness.

4 CONCLUSION

In this paper, we propose to consider fair summarization from readers' and authors' perspectives. We propose an author satisfaction-based novel evaluation measure (CROSSEM) which considers the author side. We further propose CMRA, a linear combination of ROUGE and CROSSEM. To ensure individual fairness i.e., equal distribution among authors and summarizers, we rely on inequality based index measure. In future, we plan to investigate more nuanced and efficient methods of capturing authors' preferences.

REFERENCES

[1] 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).

[2] 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).

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

[4] 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).

[5] 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*.

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

[7] 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*.

[8] 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).

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

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

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

[12] 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).

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

[14] 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).

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

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

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

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

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

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

[21] 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.

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

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

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

[25] 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)*.

[26] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* (1938).

[27] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*.

[28] Annie Louis and Ani Nenkova. 2009. Automatically Evaluating Content Selection in Summarization without Human Models. In *EMNLP*.

[29] Annie Louis and Ani Nenkova. 2013. Automatically Assessing Machine Summary Content Without a Gold Standard. In *COLING*.

[30] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* (1958).

[31] S. Mackie, R. McCreadie, C. Macdonald, and I. Ounis. 2014. Comparing Algorithms for Microblog Summarisation. In *CLEF*.

[32] Aadi Swadipto Mondal, Rakesh Bal, Sayan Sinha, and Gourab K Patro. 2021. Two-Sided Fairness in Non-Personalised Recommendations (Student Abstract). In *AAAI*.

[33] 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).

[34] 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*.

[35] 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*.

[36] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network Based Sequence Model for Extractive Summarization of Documents. In *AAAI*.

[37] Ani Nenkova and Lucy Vanderwende. 2005. *The impact of frequency on summarization*. Technical Report. Microsoft Research.

[38] 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.

[39] 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*.

[40] Gourab K. Patro, Abhijnan Chakraborty, Niloy Ganguly, and Krishna Gummadi. 2020. Incremental Fairness in Two-Sided Market Platforms: On Smoothly Updating Recommendations. In *AAAI*.

[41] 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*.

[42] Horacio Saggion, Juan-Manuel Torres-Moreno, Iria da Cunha, and Eric SanJuan. 2010. Multilingual Summarization Evaluation without Human Models. In *COLING*.

[43] 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*.

[44] 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*.

[45] B. Sharifi, M. Hutton, and J. K. Kalita. 2010. Experiments in Microblog Summarization. In *IEEE Conference on Social Computing*.

[46] C. Spearman. 1987. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* (1987).

[47] 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).

[48] 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*.

[49] 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)*.

[50] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.

[51] 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*.