Do Extractive Summarization Algorithms Amplify Lexical Bias in News Articles?

Rei Shimizu
Waseda University
Tokyo, Japan
rei-shimizu@akane.waseda.jp

Sumio Fujita
Yahoo Japan Corporation
Tokyo, Japan
sufujita@yahoo-corp.jp

Tetsuya Sakai
Waseda University
Tokyo, Japan
tetsuyasakai@acm.org

ABSTRACT

Users who read news summaries on search engine result pages and social media may not access the original news articles. Hence, if the summaries are automatically generated, it is vital that the automatic summaries represent the contents of the original articles accurately and fairly. The present study is concerned with lexical bias in sentences: a sentence is considered lexically biased if it contains expressions that may strongly influence the reader's opinion about a topic either positively or negatively. More specifically, we are interested in whether extractive summarizers can amplify lexical bias, by excessively extracting lexically biased sentences from the original article and thus misrepresent it. To address this question, we first introduce the Bias Independence Principle (BIP), which says that the probability that a sentence is selected by an extractive summarizer should be independent of whether the sentence is lexically biased or not. Based on the BIP, we propose an evaluation measure for extractive summarizers called the Bias Independence Criterion (BIC), which compares the distribution of the sentence scores for lexically biased sentences and that of the sentence scores for non-biased sentences. Moreover, based on the BIC, we define another measure called the Summary Feature Permutation Importance (SFPI) to examine whether a particular feature used by a feature-based extractive summarizer is responsible for amplifying lexical bias. Our experimental results suggest that a) Different extractive summarizers can amplify lexical bias to different degrees; b) The features useful for extracting informative sentences may also be responsible for amplifying lexical bias; and c) as mean ROUGE scores increase (implying higher informativeness), mean BIC scores also tend to increase (implying a higher concentration of lexically biased sentences).

CCS CONCEPTS

• **Information systems → Summarization**.

KEYWORDS

evaluation; evaluation measures; lexical bias; text summarization

ICTIR '22, July 11–12, 2022, Madrid, Spain

ACM ISBN 978-1-4503-9412-3/22/07…$15.00
https://doi.org/10.1145/3539813.3545123

ACM Reference Format:
Rei Shimizu, Sumio Fujita, and Tetsuya Sakai. 2022. Do Extractive Summarization Algorithms Amplify Lexical Bias in News Articles?. In *Proceedings of the 2022 ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR '22), July 11–12, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539813.3545123

1 INTRODUCTION

Text summarization and information retrieval (IR) often complement each other to enable effective and efficient information access. For example, automatically generated summaries have been used for document indexing for high-precision retrieval of highly relevant documents [22]; today, summaries (or *snippets*) are an integral component of web search result pages [24], and some IR evaluation methods take this into account (e.g., [21, 23, 25]). The present study presents a new perspective on the evaluation of text summarization where the target documents are news articles.

Users of search engines and social media often read the news summaries that they are presented with, without ever accessing the original news articles. Hence, if the summaries are automatically generated, it is vital that the automatic summaries represent the contents of the original articles accurately and fairly. More specifically, while it is natural that a news article may contain some personal views of the writer about the event being reported, if an automatic extractive summarizer selects too many sentences that are "biased" from the article, the resultant summary may mislead the reader and the misinterpretation may be further propagated on social media. Consider the following sentence: *Obama campaign spokeswoman Lis Smith described the new Romney-Ryan ad on the subject as "dishonest and hypocritical", considering Ryan's own proposals for Medicare.* In the BASIL (Bias Annotation Spans on the Informational Level) dataset [10], this sentence is labeled as containing *lexical bias*, due to the use of the expression "*dishonest and hypocritical*." In general, if a sentence contains polarized expressions that may strongly influence the reader's opinion about a topic either positively or negatively, we say that the sentence is lexically biased [10].

In the present study, we address the question of whether existing extractive summarizers amplify lexical bias by excessively extracting lexically biased sentences from the original article and thus misrepresent it. To address this question, we first introduce the Bias Independence Principle (BIP), which says that the probability that a sentence is selected by an extractive summarizer should be independent of whether the sentences is lexically biased or not. Based on the BIP, we propose an evaluation measure for extractive summarizers called the Bias Independence Criterion (BIC), which compares the distribution of the sentences scores for lexically biased sentences and that of the sentences scores for non-biased sentences. BIC is

based on an *ordinal quantification* measure called the *Normalized Match Distance*, which is a normalized version of *Earth Mover's Distance* for probability mass functions [20]. Moreover, based on the BIC, we define another measure called the Summary Feature Permutation Importance (SFPI) to examine whether a particular feature used by a feature-based extractive summarizer is responsible for amplifying lexical bias. Our experimental results suggest that a) Different extractive summarizers can amplify lexical bias to different degrees; b) The features useful for extracting informative sentences, namely, *sentence position*, *number of keywords* in the sentence, and *sentence length*, may also be responsible for amplifying lexical bias; and c) as mean ROUGE scores increase (implying higher informativeness), mean BIC scores also tend to increase (implying a higher concentration of lexically biased sentences). Hence, while summary informativeness measures such as ROUGE are widely used, our measures based on lexical bias provide a new angle to summarization evaluation if a lexical bias label is available for each sentence.

2 RELATED WORK

2.1 Dealing with Media Bias

Media bias is introduced by news media or news reporters when they strongly support a particular point of view regarding a topic. Such biases can sometimes have undesirable effects [3]. Lexical bias can be considered as a means of introducing media bias. For this reason, various studies have been conducted from the perspective of information science to analyze and detect media bias. Such studies include the investigation of media bias from a political point of view [5] [6] [13] or media bias that may compromise the trustworthiness of the news [2] [19].

2.2 Bias in News Summaries

The tendency for important sentences to be located at the beginning of sentences in news articles is called *lead bias*. There are some studies to address lead bias on automatic news summarization. Grenander et al. pointed out the problem caused by the lead bias in extractive news summarization and proposed a way to address this [11]. Zhu et al. proposed a way to deal with the effects of lead bias in abstractive news summarization [26]. However, the relationship between lead bias and media bias has not been clarified in previous work. To the best of our knowledge, our work is the first to analyze media bias in single-document extractive news summarization.

3 BIAS INDEPENDENCE PRINCIPLE

We regard the problem of extractive summarization as a sentence scoring problem (plus thresholding by scores), where each sentence is either lexically biased or not lexically biased. We posit that one possible desirable property of an extractive summarizer is to satisfy the Bias Indepence Principle (BIP), as described below.

Let $D \in \mathcal{D}$ be a document, and let $s \in D$ be a sentence. Let $l(s)(\in \{b, n\})$ be the label which indicates whether the sentence is biased or not, where b means "lexically biased" and n means "not lexically biased." Let $P_C(s \mid l(s) = X)$ be the probability that the summarizer C will include s in the summary, given that the label for the sentence is $X(\in \{b, n\})$. Then we say that C satisfies the BIC if, for any sentence s in any document D,

$$P_C(s \mid l(s) = n) = P_C(s \mid l(s) = b) \tag{1}$$

If $P_C(s \mid l(s) = b)$ is higher than $P_C(s \mid l(s) = n)$ on average, that suggests that the summarizer C tends to favor lexically biased sentences.

4 PROPOSED MEASURES

4.1 MBIC: A Measure for Quantifying Lexical Bias Amplification

Based on the BIP, we propose an evaluation measure for extractive summarizers to quantify lexical bias amplification. Given a document D, let $D_B = \{s \in D \mid l(s) = b\}$ and $D_N = \{s \in D \mid l(s) = n\}$. That is, D_B is the set of lexically biased sentences of D, and $D_N = D - D_B$. For an extractive summarizer C, let $p_C(D_B)$ and $p_C(D_N)$ denote the probability mass functions of normalized sentences scores for D_B and D_N, respectively. While the normalized sentence scores are typically continuous and constitute probability density functions, we convert them into probability mass functions over $I = 20$ bins to simplify calculations, where the first bin represents $0 \leq score(s) < 0.05$, the second bin represents $0.05 \leq score(s) < 0.1$, and so on. Let $cp_C(D_B)$ and $cp_C(D_N)$ denote the corresponding *cumulative* distributions, and let $cp^i_C(\bullet)$ denote the cumulative probability for the i-th bin. To quantify how $p_C(D_B)$ and $p_C(D_N)$ over the ordinal bins differ, we use *Normalized Match Distance* (NMD) [20] as follows:

$$NMD(p_C(D_B), p_C(D_N)) = \frac{\sum_{i=1}^{I} |cp^i_C(D_B) - cp^i_C(D_N)|}{I - 1} \tag{2}$$

Let $\mu_{p(\bullet)}$ be the mean of the probability mass function $p(\bullet)$. Let $d_C(D_B, D_N) = 1$ if $\mu_{p_C(D_B)} \geq \mu_{p_C(D_N)}$, and $d_C(D_B, D_N) = -1$ otherwise. That is, if $d_C(D_B, D_N) > 0$, the sentence scores for lexically biased sentences in D are higher than those for non-biased sentences on average. We define BIC as follows:

$$BIC_C(D) = d_C(D_B, D_N) NMD(p_C(D_B), p_C(D_N)) \tag{3}$$

Furthermore, given a document set $\mathcal{D}$, we evaluate an extractive summarizer C using *Mean* BIC (MBIC):

$$MBIC_C(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{D \in \mathcal{D}} BIC_C(D) \tag{4}$$

A positive MBIC implies the tendency of a summarizer to favor lexically biased sentences; a negative MBIC implies its tendency to favor non-biased sentences. If MBIC is close to zero, that implies that the summarizer aligns well with the BIP on average.

4.2 SFPI: A Measure for Quantifying How a Summarization Feature Affects Lexical Bias Amplification

Based on MBIC, we propose a measure called SFPI (Summary Feature Permutation Importance) to analyze how each feature used in feature-based extractive summarizers affects lexical bias. SFPI is inspired by the *permutation importance* metric used in machine learning for identifying important features [1], and examines how the MBIC for a feature-based summarizer changes when a particular feature is taken out.

Since SFPI is designed for extractive summarizers based on feature engineering (and not end-to-end neural summarizers), we consider as input a tabular dataset $\mathcal{G}$ that contains feature values obtained from each sentence instead of the set of raw documents $\mathcal{D}$. For algorithm C, let $MBIC_C(\mathcal{G})$ denote the MBIC when dataset $\mathcal{G}$ is used. Furthermore, consider a variant of $\mathcal{G}$ in which the feature values of the column representing feature f is randomly shuffled, essentially switching off the effect of f. We perform the shuffling k ($k = 1, \ldots, K$) times and the variant from the k-th trial is denoted by $\mathcal{G}_f^k$. The SFPI of feature f for C is calculated as follows:

$$SFPI_C(\mathcal{G}, f) = MBIC_C(\mathcal{G}) - \frac{1}{K}\sum_{k=0}^{K-1} MBIC_C(\mathcal{G}_f^k) \tag{5}$$

In this equation, if $SFPI_C(\mathcal{G}, f) > 0$, MBIC decreases as a result of disabling the feature f, and we can say that the feature f is a cause of amplifying lexical bias with algorithm C. In contrast, when $SFPI_C(\mathcal{G}, f) < 0$, MBIC increases as a result of disabling feature f, and it can be said that the feature f contributes to the reduction of MBIC.

5 EXPERIMENTAL SETUP

5.1 Target Summarization Algorithms

To investigate whether and how different extractive summarization algorithms amplify lexical bias, we considered LexRank [9] and TextRank [16] as unsupervised algorithms, and SummaRuNNer [17] and BERTSum [15] as supervised algorithms. Both of these supervised algorithms are neural network-based and employs a two-layer model architecture. The first layer is the encoder, which takes sentences as input and outputs an intermediate representation of them. The encoder in SummaRuNNer consists of an RNN [17], while the encoder in BERTSum [15] is a pre-trained BERT (bidirectional encoder representations from transformers) [8]. The second layer is the summarization layer, which takes the intermediate representation of the sentence obtained by the encoder as input and outputs a binary label indicating whether it should be included in the summary based on the input representation. For SummaRuNNer, Bi-GRU is used for both the encoder and the summarization layers. BERTSum uses Bi-LSTM, which is composed of a feed-forward network or transformer as the summarization layer. SummaRuNNer was trained by 5 epochs with a batch size of 32.[1] BERTSum was trained by 50,000 steps with a maximum batch size of 3,000.[2] We chose each model as the target of experiments with the highest ROUGE-1 in training iterations. All supervised algorithms were trained with the CNN/DailyMail dataset [12] with the same train/validation/test split. As this dataset is not directly applicable to extractive summarization, we used its preprocessed version as described by Cheng et al. [7].

5.2 The BASIL dataset

To evaluate the summarization algorithms in terms of lexical bias amplification, we used the aforementioned BASIL dataset [10]. This dataset consists of news articles from Fox News, Huffington Post, and New York Times, where every sentence has a label which says whether it is lexically biased or not. Note that the summarization algorithms using BERT as the input layer are limited to a sentence input length of 512 tokens [8]. Therefore, for fairness across all summarization algorithms, we processed the input sentences from the dataset to be within 512 tokens. Evaluating neural summarizers that can process longer texts (e.g. [4]) is left for future work.

[1] https://github.com/hpzhao/SummaRuNNer
[2] https://github.com/nlpyang/PreSumm

Table 1: SFPI of each feature for the random forest summarizer.

Shuffled Feature	SFPI
Sentence position	1.0210^{-2}
Number of keywords	0.2310^{-2}
Sentence length	0.7310^{-2}

6 WHICH FEATURES ARE RESPONSIBLE FOR AMPLIFYING LEXICAL BIAS?

It is known that features such as *TF-IDF*, *number of keywords*, *sentence length* and *sentence position* are important for selecting informative sentences in extractive summarization [18]. We hypothesize that these features also affect the overall lexical bias of a summary. Because summarizers described in Section 5.1 do not use such features explicitly, we use random forest as an example of a feature-based extractive summarizer, and investigated the SFPI for the above features. As for *number of keywords*, we considered proper nouns as keywords. In this experiment, the random forest summarizer was trained to output the selection probability for each sentence, using the training setting described in Section 5.1. We let the number of trials for computing SFPI be $K = 20$ on the BASIL dataset. The results are shown in Table 1. Note that we do not calculate the SFPI of *TF-IDF*, because it is a fundamental feature for any summarization algorithm to vectorize sentences. The positive SFPI values shown in the Table 1 suggest that features that are known to be useful for extracting informative sentences may also amplify lexical bias in extractive summarization. Also, the SFPI of *sentence position* is higher than the other two. That is, *sentence position* is one primary feature responsible for amplifying lexical bias, at least for the random forest summarizer.

7 EVALUATING EXTRACTIVE SUMMARIZERS WITH MBIC

7.1 How does MBIC vary depending on the algorithm?

In this section, we compare the differences of MBIC between the summarizers described in Section 5.1. While Table 1 suggests that the features may be causing the random forest summarizer to favour lexically biased sentences, we hypothesize that similar phenomena occur for neural summarizers as well, even though they do not leverage any features explicitly. As the SFPI of *sentence position* was the highest in Table 1 and this feature is relatively easy to perturb for neural summarizers, we investigated the effect of this perturbation with the neural supervised summarizers. More specifically, we randomly shuffled the order of input sentences to perturb the sentence position feature and computed the new MBIC; we averaged the MBICs over 5 trials. The results are shown in Table 2.

Note that the "Shuffled" results for LexRank and TextRank are left blank as they are unsupervised and do not explicitly depend on the sentence position feature.

The first two rows of Table 2 show that SummaRuNNer's MBIC is substantially higher than those of the unsupervised summarizers (LexRank and TextRank) and the BERTSum variants. That is, the results suggest that SummaRuNNer amplifies lexical bias much more than the other summarizers. The bottom two rows of Table 2 show that when the input sentences are shuffled so that the sentence feature is perturbed, the MBICs go down; in particular, the MBICs of the BERTSum variants are close to zero after the perturbation, which suggests that they align well with our BIP. These results suggest that sentence position is one major cause of lexical bias amplification for neural summarizers as well.

7.2 How does MBIC correlate with ROUGE?

ROUGE is a widely-used measure for evaluating the informativeness of a summary [14]. To investigate how MBIC is related to ROUGE, we compared the relationship between ROUGE-1 and MBIC across the training iterations for BERTSum. [3]

Figure 1 visualises how the ROUGE-1 scores and MBIC scores change with the training iterations for the BERTSum variants. Recall that a high ROUGE-1 score (see the solid lines) implies an informative summary, and that a high MBIC score (see the dotted lines) implies lexical bias amplification. It can be observed that, as the training proceeds, the summarizers produce more informative summaries, but at the expense of introducing more lexical bias.

Why do supervised summarization algorithms extract biased sentences as the training proceeds? We hypothesized that the sentence position feature discussed in Section 7.1 provides an explanation. We investigated the position of the reference summary sentences in CNN/DailyMail used as train data and lexically biased sentences in BASIL dataset. As a result, we found that: (i) many reference summary sentences in CNN/DailyMail are located at the beginning of articles (See the discussion of *lead bias* in Section 2); and (ii) many lexically biased sentences in the BASIL dataset are located at the beginning of a document, as shown in Figure 2. These results suggest that the summarizers are trained to favor sentences located at the beginning of a document, *and* that such sentences tend to be lexically biased. That is, this result suggests a link between lead bias (i.e., concentration of informative sentences at the beginning of a document) and media bias (as expressed using lexically biased sentences). Among the extractive summarizers that we considered, the BERTSum variants perform best in terms of ROUGE; according to Table 2, they also perform relatively well in terms of MBIC, which is consistent with the work of Liu et al. [15] who reported that BERTSum is robust to positional bias. That is, according to our experiments, BERTSum (with any summarization layer) is the best summarizer in terms of both informativeness and alignment with the BIP.

[3] We also experimented with ROUGE-2 and ROUGE-L, but the results are omitted as the trends were very similar.

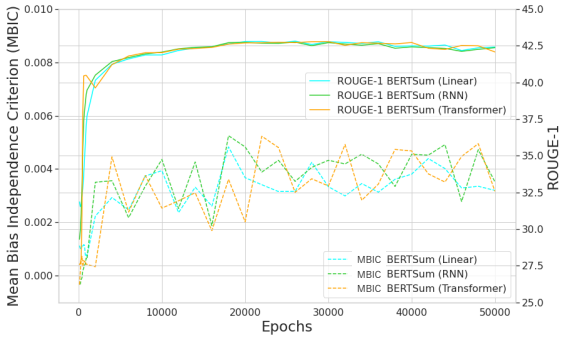

Figure 1: Relationship between MBIC and ROUGE-1 considering training iterations in BERTSum. Solid lines indicate the change of ROUGE-1. Dotted lines indicate the change of MBIC. The higher MBIC means that the algorithm tends to extract lexically biased sentences.

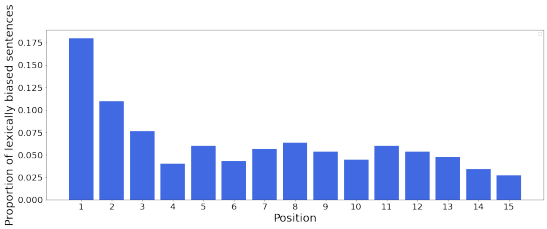

Figure 2: Distribution of lexically biased sentences over sentence positions (BASIL dataset).

8 CONCLUSION

In this paper, we presented the Bias Indepence Principle (BIP) for text summarization, and defined two evaluation measures for quantifying lexical bias amplification, namely, MBIC and SFPI. We leveraged the BASIL dataset to compute the measures for supervised and unsupervised extractive summarizers. Our conclusions are as follows.

- Different extractive summarizers can amplify lexical bias to different degrees.
- In supervised extractive summarizers, we showed that *sentence position*, *number of keywords* and *sentence length* which are useful for extracting informative sentences tend to increase MBIC. For neural summarizers, we also found that MBIC can be reduced by extracting summaries after we perturbed the *sentence position* feature, which is one of the features required for extractive summarizers.
- As mean ROUGE increases, MBIC also increases. Thus, if lexical bias labels are available, our measures can provide a new angle to summarization evaluation. Also, according to our experiments, BERTSum performs well both in terms of ROUGE and MBIC.

As future work, we would like to devise a binary classifier for automatically detecting lexically biased sentences by leveraging the BASIL dataset, so that we can apply MBIC to new data. Moreover,

Table 2: MBIC results of extractive summarizers. The rows labelled "-Shuffled" represent the results when the input sentences are shuffled to perturb the sentence position feature. For BERTSum, the network structure used in the Summarization Layer is shown in parentheses. The 95% confidence intervals are based on Student's t-distribution.

	LexRank	TextRank	SummaRuNNer	BERTSum (Linear)	BERTSum (Bi-GRU)	BERTSum (Transformer)
MBIC ($\times 10^{-2}$)	0.28	1.20	3.35	0.48	0.53	0.36
MBIC 95% CI ($\times 10^{-2}$)	[0.12, 0.54]	[0.85, 1.55]	[2.65, 4.05]	[0.25, 0.71]	[0.28, 0.77]	[0.14, 0.58]
MBIC-Shuffled ($\times 10^{-2}$)	-	-	2.44	0.01	0.13	0.16
MBIC-Shuffled 95% CI ($\times 10^{-2}$)	-	-	[1.80, 3.09]	[−0.08, 0.21]	[−0.03, 0.29]	[0.00, 0.32]

we would like to explore *summary rewriting* to reduce lexical bias in extractive summaries, and to investigate lexical bias amplification in *abstractive* summarization.

REFERENCES

[1] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. 2010. Permutation Importance: A Corrected Feature Importance Measure. *Bioinformatics* 26, 10 (2010), 1340–1347.
[2] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2020. Predicting Factuality of Reporting and Bias of News Media Sources. *Proceedings of EMNLP 2018* (2020), 3528–3539.
[3] David P. Baron. 2006. Persistent Media Bias. *Journal of Public Economics* 90, 1-2 (2006), 1–36.
[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150v2 [cs.CL]
[5] Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. 2020. Detecting Media Bias in News Articles Using Gaussian Bias Distributions. *Findings of ACL:EMNLP 2020* (2020), 4290–4300.
[6] Wei-Fan Chen, Khalid Al-Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity. arXiv:2010.10652 [cs.CL]
[7] Jianpeng Cheng and Mirella Lapata. 2016. Neural Summarization by Extracting Sentences and Words. *Proceedings of ACL 2016* 1 (2016), 484–494.
[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019* 1 (2019), 4171–4186.
[9] Güneş Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research* 22 (2004), 457–479.
[10] Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2020. In Plain Sight: Media Bias through the Lens of Factual Reporting. *Proceedings of EMNLP-IJCNLP 2019* (2020), 6343–6349.
[11] Matt Grenander, Yue Dong, Jackie C.K. Cheung, and Annie Louis. 2019. Countering the Effects of Lead Bias in News Summarization via Multi-stage Training and Auxiliary Losses. *Proceedings of EMNLP-IJCNLP 2019* (2019), 6019–6024.
[12] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. *Advances in Neural Information Processing Systems* 2015-January (2015), 1693–1701.
[13] Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. *Proceedings of ACL 2014* 1 (2014), 1113–1122.
[14] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, 74–81.
[15] Yang Liu and Mirella Lapata. 2020. Text Summarization with Pretrained Encoders. *Proceedings of EMNLP-IJCNLP 2019* (2020), 3730–3740.
[16] Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into Texts. *Proceedings of EMNLP 2004* (2004), 404–411.
[17] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents. *Proceedings of AAAI 2017* (2017), 3075–3081.
[18] Anshuman Pattanaik, Sanjeevani Subhadra Mishra, and Madhabananda Das. 2020. A Comparative Study of Classifiers for Extractive Text Summarization. *Advances in Intelligent Systems and Computing* 1101 (2020), 173–181.
[19] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-checking. *Proceedings of EMNLP 2017* (2017), 2931–2937.
[20] Tetsuya Sakai. 2021. Evaluating Evaluation Measures for Ordinal Classification and Ordinal Quantification. *Proceedings of ACL-IJCNLP 2021* (2021), 2759–2769.
[21] Tetsuya Sakai and Zhicheng Dou. 2013. Summaries, Ranked Retrieval and Sessions: A Unified Framework for Information Access Evaluation. In *Proceedings of ACM SIGIR 2013*. 473–482.
[22] Tetsuya Sakai and Karen Sparck Jones. 2001. Generic Summaries for Indexing in Information Retrieval. In *Proceedings of ACM SIGIR 2001*. 190–198.
[23] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-based Calibration of Effectiveness Measures. In *Proceedings of ACM SIGIR 2012*. 95–104.
[24] Anastasios Tombros and Mark Sanderson. 1998. Advantages of Query Biased Summaries in Information Retrieval. In *Proceedings of ACM SIGIR '98*. 2–10.
[25] Andrew Turpin, Falk Scholer, Kalervo Järvelin, Mingfang Wu, and J. Shane Culpepper. 2009. Including Summaries in System Evaluation. In *Proceedings of ACM SIGIR 2009*. 508–515.
[26] Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging Lead Bias for Zero-shot Abstractive News Summarization. *Proceedings of ACM SIGIR 2021* (2021), 1462–1471.