Fairness of Extractive Text Summarization

Anurag Shandilya
IIT Kharagpur, India

Kripabandhu Ghosh
IIT Kanpur, India

Saptarshi Ghosh
IIT Kharagpur, India

ABSTRACT

We propose to evaluate extractive summarization algorithms from a completely new perspective. Considering that an extractive summarization algorithm selects a subset of the textual units in the input data for inclusion in the summary, we investigate *whether this selection is fair*. We use several summarization algorithms over datasets that have a sensitive attribute (e.g., gender, political leaning) associated with the textual units, and find that the generated summaries often have very different distributions of the said attribute. Specifically, some classes of the textual units are under-represented in the summaries according to the fairness notion of *adverse impact*. To our knowledge, this is the first work on fairness of summarization, and is likely to open up interesting research problems.

CCS CONCEPTS

• **Information systems → Information retrieval**;

KEYWORDS

Extractive summarization; fairness; adverse impact

ACM Reference Format:
Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of Extractive Text Summarization. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3184558.3186947

1 INTRODUCTION

With the phenomenal increase in the amount of textual information on the Web, text summarization algorithms are commonly used to get a quick overview of the information. Most existing summarization algorithms are *extractive* in nature (as opposed to abstractive). Extractive algorithms form the summary by extracting some of the textual units in the input, e.g., individual sentences in a document, or individual microblogs in a set of microblogs. Traditionally, summarization algorithms are judged on how closely the algorithmic summary matches gold standard summaries (usually written by human annotators), using measures such as ROUGE scores. In this work, we propose to look at extractive summarization algorithms from a completely new perspective.

Extractive summarization algorithms essentially perform a selection of a (small) subset of the textual units in the input, for inclusion in the summary, based on some measure of the relative quality or importance of the textual units. We propose to investigate *whether this selection is fair*, i.e., whether the generated summary is a fair representation of the input data.

WWW '18 Companion, April 23–27, 2018, Lyon, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186947

An algorithm is said to be unfair if it denies a desired outcome to an individual or a group of individuals on grounds that are inappropriate, e.g., based on a protected attribute according to laws of a country. In the context of summarization, we assume that the textual units in the input data (sentences, tweets, etc.) are associated with a sensitive or protected attribute, such as political leaning of the text or the gender/race/religion of the person who has written the text. In this context, the desired outcome is to be included in the summary, assuming that only the textual units included in the summary will get visibility or be seen by humans (and not the entire input data). Our motivation for investigating the fairness of summarization algorithms comes from the concern that using a (inadvertently) 'biased' summarization algorithm can further reduce the visibility of the voice / opinion of a certain group in the summary.

Let the distribution of the sensitive/protected attribute in the input data be termed as the 'input distribution'. The summary will contain a subset of the textual units, which will also have a distribution of the same sensitive attribute; let this distribution be the 'summary distribution'. We propose to judge the fairness of a summarization algorithm by comparing the input distribution and the summary distribution, based on fairness notions such as 'adverse impact' of the U.S. Equal Employment Opportunity Commission [1].

We perform experiments with two datasets of microblogs/tweets, one which contain tweets written by users having known gender, and the other containing tweets having known political leaning. We apply several well-known summarization algorithms over these datasets. We find that the distribution of the sensitive attribute (gender or political leaning) is very different in the summaries generated by different algorithms. In fact, some specific classes of the textual units (tweets) are *under-represented* according to the 'adverse impact' notion [1] in the summaries generated by several algorithms. To our knowledge, this is the first study which demonstrates that applications of summarization algorithms need to consider the fairness of such algorithms, along with the standard measures of the quality of the summaries.

2 EXPERIMENTS AND OBSERVATIONS

We now describe the datasets and extractive summarization algorithms considered for the present work.

Datasets: We considered two datasets in which every textual unit is annotated with a sensitive attribute (gender or political leaning).
(1) Claritin dataset, that contains tweets about the effects of the drug Claritin (details at *https://tinyurl.com/claritindataset*). Each tweet is annotated with the gender of the user who posted it. After removing exact duplicates, we have 4,111 tweets in total, out of which 1,556 (37.8%) are written by male users, while 2,555 (62.2%) are written by female users.
(2) US Election dataset [2], that contains tweets posted during the 2016 US Presidential elections. Each tweet is annotated as supporting or attacking one of the presidential candidates (Donald Trump

and Hillary Clinton) or neutral or attacking both. For simplicity, we grouped the tweets into three classes: (i) *Pro-Republican*: tweets which support Trump and / or attack Clinton, (ii) *Pro-Democratic*: tweets which support Clinton and / or attack Trump, and (iii) *Neutral*: tweets which are neutral or attack both candidates. After removing exact duplicates, we have 3,450 tweets in total, out of which 2,115 (61.3%) are Pro-Republican, 1,114 (32.3%) are Pro-Democratic, and 221 (6.4%) are Neutral tweets.

Summarization algorithms: We consider seven well-known extractive summarization algorithms. These algorithms generally estimate an importance score for each textual unit (sentence / tweet) in the input, and include the textual units in the summary in decreasing order of this score, until a pre-defined length of the summary is reached. **(1) Cluster-rank [5]** 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** [7] which measures the relationship between the textual units using linear combinations and reconstructions, and generates the summary by minimizing the reconstruction error; **(3) Frequency Summarizer** [4], which assumes that if a textual unit contains the most frequent words, it is likely to be a good candidate for including in the summary; **(4) LexRank** [3], which computes the importance of textual units using eigenvector centrality on a graph representation based on similarity of the units; **(5) LSA** [6], which constructs a terms-by-units matrix, and estimates the importance of the textual units based on SVD on the matrix; **(6) LUHN** [8], which derives a 'significance factor' for each textual unit based on occurrences and placements of frequent words within the unit; **(7) SumBasic** [9], which uses frequency-based selection of textual units, and re-weighting of the word probabilities to minimize redundancy.

Results: We apply the summarization algorithms stated earlier on the two datasets, to obtain summaries of length 100 tweets each. Table 1 shows the results of summarizing the two datasets – shown are the numbers of tweets of the different classes in the whole dataset (first row), and in the summaries generated by the different summarization algorithms (subsequent rows).

It is clear that summaries produced by different summarization algorithms vary considerably, and contain very different distributions of the gender / political leaning, as compared to the distribution in the whole input data. For instance, for the Claritin dataset, the LSA algorithm increases the proportion of tweets from female users in the summary, while FreqSum, DSDR, LexRank and LUHN increase the proportion of tweets from male users in their summaries. In fact, while the whole dataset contains pre-dominantly tweets from female users (only 37.8% from male), the summaries by LexRank and LUHN contain pre-dominantly tweets from male users (51% and 55% male). The observations are similar for the US Election dataset. While the whole dataset is predominantly pro-Republican (61.3%), all the algorithms reduce the proportion of pro-Republican tweets in the summary. Especially, the summaries generated by LexRank and LSA have the opposite majority, i.e., include more pro-Democratic tweets than pro-Republican tweets. Clearly, a person who reads the summaries might form a very different view or opinion, as compared to reading the whole dataset.

We verify the fairness of the summarization using the notion of 'adverse impact' used by the U.S. Equal Employment Opportunity Commission to determine whether a company's hiring policy has any adverse impact on a demographic group [1]. According to this policy, a particular class c is *under-represented* (disadvantaged) in the selected set (i.e., the summary), if the fraction of selected items belonging to class c is less than 80% of the fraction of selected items of that class which has the highest selection rate. Applying this rule, we find under-representation of particular classes of tweets in the summaries generated by many of the algorithms; these cases are marked with an asterisk (*) in Table 1.

Method	Claritin dataset		US Election dataset		
	Male	Female	Pro-Rep	Pro-Dem	Neutral
Whole data	1,556 (37.8%)	2,555 (62.2%)	2,115 (61.3%)	1,114 (32.3%)	221 (6.4%)
ClusterRank	37	63	55*	38	7
DSDR	43	57	56*	38	6*
FreqSumm	46	54*	52*	43	5*
LexRank	51	49*	48*	49	3*
LSA	32*	68	42*	46*	12
LUHN	55	45*	58	35	7
SumBasic	37	63	52*	44	4*

Table 1: Results of summarizing the (i) Claritin dataset, and (ii) US Election dataset: Number of tweets of different classes, in the whole dataset and the summaries of length100 tweets computed by different algorithms. * indicates under-representation of a class according to the fairness notion of 'adverse impact' [1] (details in text).

We repeated the experiments for summaries of lengths other than 100 as well, such as for $50, 200, 300, \ldots, 500$ (details omitted due to lack of space). We observed several cases where the same algorithm includes very different proportions of tweets of various classes, while generating summaries of different lengths. Hence, whether summarization is fair depends on several factors, including the particular algorithm used and the length of summary.

3 CONCLUSION AND FUTURE DIRECTIONS

This is the first work that introduces the concept of fairness of text summarization algorithms. We show that while summarizing datasets having an associated sensitive attribute, one needs to verify the fairness of the summary. Especially, with the advent of neural network-based summarization algorithms (which involve supervised learning), the question of fairness becomes even more critical. We believe that this work will open up interesting research problems, e.g., on developing algorithms that will ensure some degree of fairness in the summaries.

REFERENCES

[1] Dan Biddle. 2006. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing.* Routledge.
[2] K. Darwish, W. Magdy, and Zanouda T. 2017. Trump vs. Hillary: What Went Viral During the 2016 US Presidential Election. In *Proc. Social Informatics (SocInfo).*
[3] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based Lexical Centrality As Salience in Text Summarization. *J. Artif. Int. Res.* 22, 1 (2004).
[4] freqsum 2014. Text summarization with NLTK. (2014). https://tinyurl.com/frequency-summarizer.
[5] Nikhil Garg and others. 2009. Clusterrank: a graph based method for meeting summarization. In *Proc. INTERSPEECH.* ISCA.
[6] Yihong Gong and Xin Liu. 2001. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis. In *Proc. ACM SIGIR.*
[7] Zhanying He and others. 2012. Document Summarization Based on Data Reconstruction. In *Proc. AAAI Conference on Artificial Intelligence.*
[8] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM J. Res. Dev.* 2, 2 (April 1958), 159–165.
[9] Ani Nenkova and Lucy Vanderwende. 2005. *The impact of frequency on summarization.* Technical Report. Microsoft Research.