

Towards Detecting Media Bias by Utilizing User Comments

Sevgi Yigit-Sert
Ankara University
Ankara, Turkey
syigit@ankara.edu.tr

Ismail Sengor Altingovde
Middle East Technical University
Ankara, Turkey
altingovde@ceng.metu.edu.tr

Özgür Ulusoy
Bilkent University
Ankara, Turkey
oulusoy@cs.bilkent.edu.tr

ABSTRACT

Automatic detection of media bias is an important and challenging problem. We propose to leverage user comments along with the content of the online news articles to automatically identify the latent aspects of a given news topic, as a first step of detecting the news resources that are biased towards a particular subset of such aspects.

CCS Concepts

•Information systems → Document topic models;

1. INTRODUCTION

Bias can exist in various stages of the publishing process, as in the selection of stories, depth of attention given to a story, or the reporting of the story [7]. While media bias is a well-explored topic in social sciences, its detection is very hard even for the humans due to the inherent subjectivity involved; and developing automatic techniques for this purpose is a recent and challenging research direction.

To detect media bias automatically, we envision a system that clusters the *news resources* on a *per-topic* basis, such that those newspapers that essentially cover the same *aspects* of a given news topic can be grouped together. In this work, as a first step towards this goal, we focus on the task of detecting the latent aspects of the news topics. To this end, we propose to leverage the user comments in addition to the content of the articles, as the latter can be a useful yet inadequate source of information on its own to detect alternative aspects. This is due to the observation that most of the articles on a particular topic, even if published at different news outlets, would overlap in the majority of the reported information (e.g., see [6]). In contrast, user comments are rather informal yet more explicit than the article content and thus, may serve as a better resource to infer the aspects of a news topic (especially for those aspects that are *subtly* emphasized by the article, in case of a bias).

This paper investigates the performance of the Latent Dirichlet Allocation (LDA) [1] using either news articles or

user comments, or both, to detect the aspects of a given news topic. In our experiments, we use a novel dataset of 2,193 news articles and 3,923 comments on two highly debated political topics in Turkey; and show that utilizing both the news and their comments yields promising results.

Related Work. Quoting patterns and appearance frequency of a political party or its members are widely used to characterize the bias in mainstream and/or social media [3, 4, 2]. In [8], topics inferred from news and tweets are compared to detect bias. Park et al. [5] apply sentiment detection approaches to label each commenter with a political ideology and then propagate these labels to the news articles. In contrast to these works, as a first step towards detecting bias, we employ the article content and user comments together to infer the aspects of news topics.

2. DATASET AND METHODOLOGY

Dataset. Our dataset includes the news articles and corresponding user comments for two highly polarizing topics in Turkish political arena. For both topics, there are two major aspects that are supporting or against the government's view, as well as several other minor aspects. For instance, for the news topic "interest rates", two major aspects argued in the news articles and/or posted comments are defending an increase in the central bank interest rates, or just the opposite; while minor aspects may involve, say, the relationships to finance policy, foreign policy, etc. For brevity, we refer to these topics as Topic-1 and Topic-2. We collected the news articles and user comments for our topics from the web portals of 6 daily newspapers in Turkey, namely, Bugün, Cumhuriyet, Sözcü, Sabah, Takvim, and Yeni Akit (referred to as resources R1 to R6, respectively), that are known to represent different political tendencies.

Annotation study. To construct the ground truth (GT) for the aspects, we collected the instant query suggestions from Google for each topic, as these suggestions are likely to be based on the popular queries and hence reflect the possible different interpretations/aspects of a topic. Then, we post-processed these suggestions, i.e., merged those that refer to the same notion, or discarded those that refer to irrelevant issues. This yielded five aspects for the first topic and four aspects for the second. Since the actual aspect descriptions would require background information in Turkish politics and clutter the text, we simply refer to these aspects with capital letters, from A to E.

We labeled each news article and comment with only one (the most dominant) aspect; or with N if none of the aspects seem relevant. For the first topic, which has a large num-

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

WebSci '16 May 22-25, 2016, Hannover, Germany

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4208-7/16/05.

DOI: <http://dx.doi.org/10.1145/2908131.2908186>

Table 1: The number of the news articles and comments from each resource for each topic aspect

		Topic-1					Topic-2					
		A	B	C	D	E	N	A	B	C	D	N
R1	News	172	27	3	6	12	81	72	38	9	1	91
	Comm.	132	29	3	31	6	98	18	13	1	0	81
R2	News	76	58	12	29	7	25	3	14	6	17	11
	Comm.	35	66	1	29	4	154	0	0	0	6	4
R3	News	30	8	1	17	6	15	7	13	2	15	16
	Comm.	58	16	0	41	0	182	3	1	145	232	288
R4	News	10	50	21	2	1	30	2	22	4	0	10
	Comm.	3	36	0	0	0	5	7	31	2	0	45
R5	News	17	113	10	6	0	37	1	53	11	0	25
	Comm.	20	175	5	6	1	98	8	34	25	0	170
R6	News	1	150	6	4	2	35	0	4	4	0	7
	Comm.	6	139	4	4	1	112	0	0	0	0	6

ber of related news/comments, the annotation covered 1,080 articles and 1,500 comments (chosen uniformly at random); while for the second topic, we annotated all the available data. Table 1 shows that for the majority of the cases, the news articles are relevant to one of the main contradicting aspects (i.e., A or B), while comments span a larger spectrum of major and minor aspects, which might be expected. Furthermore, a much larger percentage of comments are labeled with N, implying that a *topic drift* is more likely for the comments during the course of discussion among the users.

Methodology. The preprocessing of the data involved conversion to lowercase, tokenization, stop-word removal and stemming of the terms that appear in the articles and comments. We used the LDA algorithm [1] (with Gibbs sampling) as provided in the Mallet library¹. In our setup, the word “topic” refers to the *news topics*; so to avoid confusion, we refer to the LDA output as the *inferred aspects*.

We model each newspaper as a single virtual document, which is composed of either all the news, comments or both, for a given news topic. For every news topic, we feed these documents to the LDA algorithm to discover its aspects. The number of aspects, k , is also given as an input to LDA. We use the most probable 7 terms to represent an aspect.

3. EXPERIMENTS AND DISCUSSIONS

In our experiments, we aim to identify the most useful content type (i.e., article text, comments or both) to detect the aspects of news topics. In Table 2, we report the results for the best-performing cases, i.e., for the k value that yielded the aspects that are semantically closest to the ground truth annotation, when only article texts are utilized. For Topic-1, the most meaningful result is obtained for $k=3$. In this case, the inferred aspect with the highest weight includes only the generic terms about the topic (such as “interest”, “rate”, “economy”, etc.) without any specific reference to its actual aspects. The other two inferred aspects are representing ground-truth aspects A and B, but with a weight that is much less than the actual (cf. see Table 1). For the second topic, which is simpler, the situation is better as both of the major aspects can be discovered. Our findings imply that while being useful, the articles may fail to yield the most-polarizing aspects of a topic, as their content include lots of generic/neutral terms about the topic.

¹<http://www.cs.umass.edu/~mccallum/mallet>

Table 2: The LDA aspects inferred using the article content (along with the closest GT aspect)

	Topic-1			Topic-2		
	No.	Weight	GT Aspect	No.	Weight	GT Aspect
1	0.266	A	1	0.802	A	
2	0.022	B	2	0.198	B	
3	0.712	neutral				

Table 3: The LDA aspects inferred using both the article content and comments (along with the closest GT aspect)

	Topic-1			Topic-2		
	No.	Weight	GT Aspect	No.	Weight	GT Aspect
1	0.829	A	1	0.919	A	
2	0.167	B	2	0.081	B	
3	0.004	A				

When LDA is applied to user comments, we find that the inferred aspects are almost never related the actual aspects of Table 1, as the representative terms obtained from comments are quite diverse and rather irrelevant to the actual topic in most of the cases. We think that the failure of user comments for discovering aspects is caused by the topic drift in comment threads, as users often discuss earlier similar news and/or involve in a discussion on a subject that is more general than the news topic at hand.

Next, we apply LDA on both the article content and user comments. In this case, best aspects are again found when k is set to 3 and 2, for the Topic-1 and Topic-2, respectively. The results reported in Table 3 show that the major aspects A and B are identified more accurately in this case (e.g., the neutral aspect is eliminated for Topic-1). A comparison of the actual terms describing each inferred aspect also revealed that more specific terms could be identified for each aspect, in comparison to the case of solely using articles.

We conclude that utilizing both the article content and associated comments is more promising to discover the aspects of a given news topic, especially when there is a large number of available comments. In our future work, we plan to exploit such aspects to detect and cluster biased newspapers, which are very likely to emphasize a certain subset of aspects, and overlook others.

4. REFERENCES

- [1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- [2] A. Dallmann, F. Lemmerich, D. Zoller, and A. Hotho. Media bias in german online newspapers. In *HT '15*, pages 133–137, 2015.
- [3] Y.-R. Lin, J. P. Bagrow, and D. Lazer. Quantifying bias in social and mainstream media. *SIGWEB Newsl.*, pages 5:1–5:6, 2012.
- [4] V. Niculae, C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, and J. Leskovec. Quotus: The structure of political media coverage as revealed by quoting patterns. In *WWW '15*, pages 798–808, 2015.
- [5] S. Park, M. Ko, J. Kim, Y. Liu, and J. Song. The politics of comments: predicting political orientation of news stories with commenters’ sentiment patterns. In *CSCW '11*, pages 113–122, 2011.
- [6] S. Park, S. Lee, and J. Song. Aspect-level news browsing: Understanding news events from multiple viewpoints. In *IUI '10*, pages 41–50, 2010.
- [7] D. Saez-Trumper, C. Castillo, and M. Lalmas. Social media news communities: Gatekeeping, coverage, and statement bias. In *CIKM '13*, pages 1679–1684, 2013.
- [8] A. Younus, M. A. Qureshi, S. K. Kingrani, M. Saeed, N. Touheed, C. O’Riordan, and G. Pasi. Investigating bias in traditional media through social media. In *WWW '12*, pages 643–644, 2012.