Multidimensional Analysis of Gender and Age Differences in News Consumption

Jisun An
Qatar Computing Research Institute,
Hamad Bin Khalifa University
Doha, Qatar
jan@qf.org.qa

Haewoon Kwak
Qatar Computing Research Institute,
Hamad Bin Khalifa University
Doha, Qatar
hkwak@qf.org.qa

ABSTRACT

Using 103,133 news items collected from Daum News, the second most popular news portal in South Korea, we provide multidimensional analyses of gender and age differences in news consumption. We quantify such differences in four different dimensions: 1) by actual news items; 2) by topic 3) by issue; and 4) by angle. We found that angle preferences can better explain the difference of news consumptions, not topics nor issues.

1. INTRODUCTION

Demographics plays an important role in news consumption. What women in their fifties read would be very different from that of men in their twenties. Understanding the differences in news consumption can potentially help journalists to pitch news articles better, help editors to decide which ones to put in the front page, or help computer scientists to design a new algorithm for recommending news articles. That is the reason why news consumption of demographic groups has been actively studied in both domains: the study of journalism and computer science.

Scholars of the study of journalism have long examined how demographics, such as sex or age, relates to information-seeking behavior [4] and news preferences. When attending to news, the sexes typically pursue remarkably different interests–in terms of topics, men tend to follow the news on politics, sports, and business and finance, whereas women turn to news about community and health issues [9, 5]. Also, women read more about social/interpersonal issues than men, and men read more about achievement/ performance than women [6]. While topic preferences have been extensively studied, little is known about the issue or angle preferences of different demographic groups' news consumption. The key reason for this oversight is mainly due to a lack of data.

On the other hand, since news sites have been publishing online, we now have access to large-scale data of individual news consumption with detailed personal profiles. Computer scientists have been focusing on building a better news recommending system to give readers a personalized experience when reading the news. Systems that make recommendations according to demographic classes were initially introduced [8]. More recently, the demographic information has often been used as a feature [3, 7].

In this study, we attempt to bridge those two worlds and uncover the differences in news consumption across demographic groups using large-scale news consumption data. Specifically, we aim to quantify such differences in four dimensions: actual news item, topic, issue, and angle. While the existence of the "differences" is expected, our multidimensional analysis shows how such differences can be differently captured in each dimension.

We collected and analyzed the top 30 news items for each gender and age group in Daum News, the second most popular news portal service in South Korea, for the entire year of 2015. The number of the news items collected is 103,133. Daum News can have the accurate, not self-reported, information on the user's age and gender, although this may not be common in the Western web services, through one's social security number. In South Korea, to join a website, it is mandatory to provide the social security number that contains the birth year and gender. Also, Daum News has a strong user base who reads news with a logged-in status mainly because Daum News offers a wide range of services, which are e-mail, Internet community, or messenger, for example, based on logged-in status.

2. DATA

2.1 News consumption in South Korea

Online news consumption in South Korea has increased drastically. About 86% of Korean people access news online at least once a week[1]. Given 92% Internet and 85% smartphone penetrations, such a drastic increase makes sense. Web portal sites such as Naver and Daum are especially popular digital news platforms. Due to the extreme popularity of these portals, news providers in Korea have been eager to publish their content via portals for years. In 2015, Naver and Daum together formed the Committee for the Evaluation of News Partnership, complete with a set of ethical standards to help decide which providers should be eligible to supply news to portals. As a result, we can fairly say that news providers and newsreaders of Daum News are representative of the general South Korea news media and population, respectively.

[1] http://www.digitalnewsreport.org/

Computation+Journalism Symposium '16 San Francisco, California, USA

DOI: 10.1145/1235

2.2 Data Collection

Daum is the second biggest web portal in South Korea, followed by Naver. The two portals play a significant role in providing a place for accessing news to Korea; 41% of Korean (24.6M users) access Daum news weekly base.

Daum News provides different ways to explore news articles, for example, by its recency, by current issues, or by regions. It also provides a ranked list of news articles based on the number of views or the number of comments. A unique feature of Daum news is that it provides the top 30 most popular news articles on a particular day for each gender and age group, which are [male or female], and [10s, 20s, 30s, 40s, and 50s and above].

We collected the top 30 most popular articles on different age and gender groups of each day for one year period (01/01/2015−31/12/2015). In our data set, we have 103,133 listed news items with 54,274 unique news titles. For each news item, we have its unique news ID, demographic group, rank in the group, title, summary, news source, and published date. We note that news articles about entertainment and sports are not included in the lists due to Daum News's policies.

3. METHODOLOGY

We determined the topical category and issue category of the news titles we collected. Next, we will briefly describe the methodology adopted for the analysis.

3.1 Categorizing News: Topic

In order to infer the topical categories of the news articles, we used the meta information embedded in the news URLs. For example, the URL http://media.daum.net/society/labor/newsview?newsid=20160709180100906l is categorized as "Society." We parsed all of Daum News's URLs and extracted the topic information.

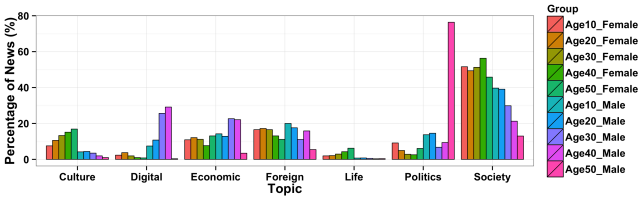


Figure 1: Group distribution by topic.

Figure 1 shows the proportion of news items for each demographic group, for each topic. Daum News has seven different news sections. The Culture, Digital, Economic, Foreign, Life, Politics, and Society sections were used, and the Entertainment and Sports sections were excluded. We observe that Female groups are alike than Male groups in terms of topical interest. For Female groups, Society and Culture were the two most popular topics. For Age10_Male and Age20_Male groups, Society, Foreign, and Politics were more popular than the other topics. Then, for the Age30_Male and Age40_Male groups, their interests are in topics like digital (i.e., technology news), economics, and politics. We also find one noticeable behavior of the Age50_Male group with having 78% of the top 30 most popular news items over a year in politics. These topical preferences align well with previous work on sex-typed news consumption internationally [9, 5].

3.2 Categorizing News: Issue

The topical category provided by Daum News abstractly captures the topic of a news item. In this study, we go beyond a mere topic preference and examine whether the gender- or age-specific issue or angle preferences exist. To do that, we need to understand the semantic content of the news articles. For example, news about "Violence at daycare" and "A killer of his family" are both categorized as society topics, but they are two different *issues*.

We automatically discover an issue-specific categorical structure from a set of titles and classify each news article based on it. Topic modeling techniques such as probabilistic Latent Semantic Indexing (pLSI), Latent Dirichlet Allocation (LDA) [1], and Hierarchical Dirichlet Processes (HDP) [10] can be employed to induce issues from the set of news titles. We manually compare the three methods in terms of the interpretability of the induced issues and the quality of the clustered news titles for each of the included issues. We find that, for our dataset, pLSI works better than LDA and HDP. Thus, for the last of our analysis, we use pLSI to detect candidate issues.

One problem with these topic modeling techniques is that they are not time sensitive. The following two news titles, "Anyang Killer – a man killed his family" and "Wife killer – a man killed his wife but did not show any grief," are likely to be categorized as the same issue even though one happened in January 2015, and the other happened in October 2015. To handle this problem, we first split our dataset by month, and then we built a pLSI model for each of monthly dataset. Each pLSI model induces 100 candidate issues, keeping us with 1,200 candidate issues in total. We then aggregate candidate issues if they have similar feature vectors.

After this, we classify each news title into one of these candidate issues based on the score given by the pLSI model. Finally, for every candidate issue, we split each further into multiple issues by considering the publication time of news items. Only news items published for days in a row are tagged as the same issue. As a result, we split 1,200 candidate issues into 2,122 issues. Those news titles without any matching candidate issue, we consider them as stand-alone issues. All together we have 41,452 issues.

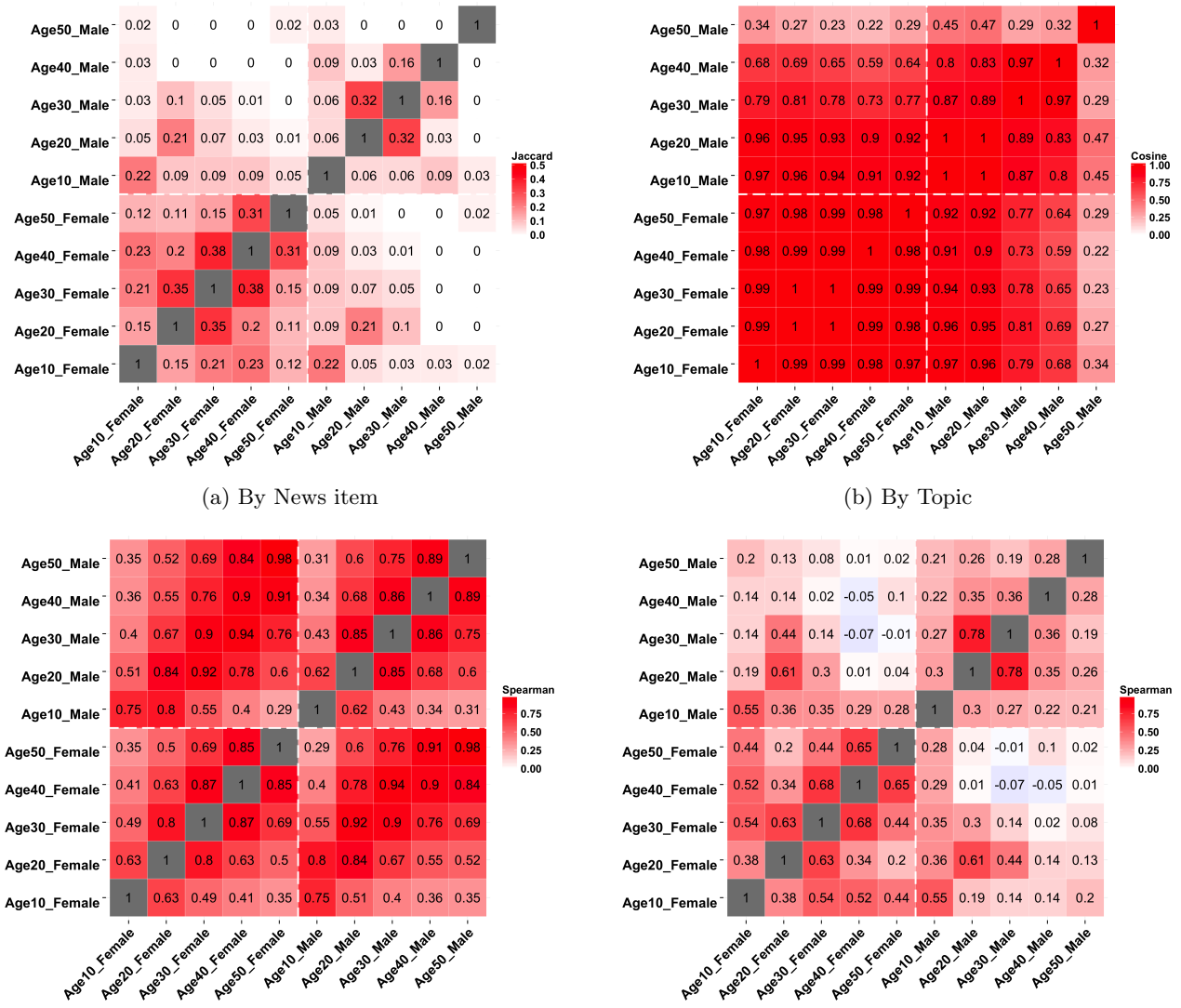
Across all groups, the most popular issue was Middle East Respiratory Syndrome (MERS) outbreak in South Korea. One issue about MERS lasts for 31 days with 2,123 articles. Across all issues, we find 126 MERS-related issues with 2,756 matching news items. The most frequent words for this issue are like MERS, a vice prime minister, confirmed patients, infected, hospital, Daejeon, tourists, etc. Such word analysis will be the base for examining angle preferences of different demographic groups. We will discuss more in Section 4.4.

4. GROUP DIFFERENCES IN NEWS CONSUMPTION

We now quantify differences in news consumption across demographic groups in four different dimensions: 1) by actual news item, 2) by topic, 3) by issue, and 4) by angle.

4.1 By news item

As a first attempt to compare news consumptions of different groups, we look into the actual news items. We measure the similarity among groups based on commonly consumed news items (by their unique news IDs) among all the top 30



(a) By News item

(b) By Topic

(c) By Issue

(d) By Angle

Figure 2: Heatmap showing the similarity across groups based on comparison of (a) news items – Jaccard similarity scores, (b) topics – Cosine similarity between topic vectors, (c) issues – Spearman's rank correlation test of two lists of news issues ranked by their lifespans, and (d) angles – Spearman's rank correlation test of two lists of words of news titles about MERS outbreak in South Korea ranked by their frequencies.

articles of each group over a one-year period. We use Jaccard Similarity to compute group similarity. For two sets A and B, the Jaccard Similarity is given by $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. For our case, let A and B be sets of news items corresponding to the two groups to be compared. Strictly, $|A \cap B|$ would translate to the count of the news items matched across the sets of A and B. Figure 2(a) shows the Jaccard similarity among groups as a heatmap. For example, the pair of Age10_Female and Age20_Female has a Jaccard score of 0.15. This means that, among the union set of all their consumed news items, 15% are common. The higher the similarity score is, the more news items are viewed in common between the two groups.

We find that, within same-sex groups, the similarity generally increases as the age difference decreases, with Female groups have a stronger tendency of it than Male groups (the average similarity score among all pairs of Female groups is 0.217 while that of Male groups is 0.071). However, we observe two exceptions, Age10_Female and Age10_Male. They are more similar to Age30 or Age40 than Age20 same-sex groups. In Figure 1, we can see that the Age10 groups have more Politics and Foreign news items in the top lists, indicating their similarity to older groups.

We then find strikingly low similarity scores between different sex groups. Age40_Male and Age50_Male have almost no news items in common with the Female groups, and the same happens for the Age40_Female and Age50_Female groups. With these results, we can conclude that the set of popular news items that females consume is very different from what males read.

4.2 By Topic

We have shown that there is a striking difference between the popular news items for the Male groups and those of

the Female groups. Now, we will examine the news consumption of those groups in terms of topical interests. This analysis will tell us whether the existing framework of news consumption based on sex or age is also found in Korea.

For each group g, we created a topic vector $T_g = (w_{1,g}, w_{2,g}, \ldots, w_{t,g})$ in which each dimension corresponded to a predefined topic from Daum News where $t = 7$ in our case (see 7 topics in Figure 1). The weight $w_{t,g}$ was computed by the proportion of the news items in the topic t for the group g. We then computed the cosine similarity between two vectors to compare the topical interests of two groups. The results are shown in Figure 2(b) as a heatmap.

For Female groups, we find high similarity scores between all pairs (> 0.97), showing that the proportion of news items in each topic are similar to each other. A similar pattern is also observed for Male groups, but to a lesser extent, and with one exception, Age50_Male, which shows very different topical interest. The reason is that they exclusively read political news – 83% of the top 30 news items for a one-year period are about politics (see Figure 1).

Male groups are further split into two groups, as Age10_Male and Age20_Male are more similar to Female groups, but Age30_Male and Age40_Male have topical interests distinct from those of Female groups. This partly supports the traditional sex-typed news consumption theory – our data set also shows different topic preferences of different gender groups. However, we find such differences are driven more by Age30 and Age40 Male groups and less by Age10 and Age20 Male groups.

In summary, the topical interests seem to be alike across all groups, except Age50_Male. Considering that news consumption largely depends on current, local issues, this could make sense. However, given the striking differences in common news items, the observation we made in the previous section (see Figure 2(a)), the fact that groups largely share topical interests, is still surprising. We now move onto the similarities in issues that different groups consume.

4.3 By Issues

Given that the topical interests are similar among groups but not the actual news items, it is intuitive to think that even if two groups are visiting the same news section such as Society, they might consume different issues – older people might read more about "Baby killer" while young people read more about "Violence at school."

To investigate such issue-specific differences in news consumption, we map each news items to a specific issue. The issues are identified by the method we described in Section 3.2. Then, we quantify the importance or the level of attention to a specific issue for a group by computing the lifespan. We define the lifespan of an issue as the longest period time where that issue appeared on the top 30 list for each group.

We then measure the similarities between groups based on the importances of the different issues. We select issues that are consumed by at least two groups, resulting in 36,134 issues. For each of these issues, we compute its lifespan for each group. This gives us a ranked list of issues for each group, and we use Spearman's rank correlation coefficient to compare two ranked lists. Figure 2(c) shows the results as a heatmap. All pairs of rankings are statistically significantly different ($p < 0.05$).

In this heatmap, we compare pairs of values. For example, a value of 0.9 between Age40_Female and Age40_Male is hard to interpret by itself. Comparing one similarity score to other entries, one observes that this value is higher to that for the 'Age40_Male' – 'Age30_Male' pair or 'Age40_Male' – 'Age50_Male' pair. Simply put, one could claim that gender differences lead to more strongly pronounced news consumption than 10 years of age difference.

By comparing pairs of values, we observe age difference plays an important role in news consumption – a similar pattern was also found when looking at common news items in Section 4.1. Given that a pair of different sex groups have so few common news items consumed, the high similarity between two ranked list of issues (Spearman's correlation coefficient $\rho > 0.8$) is striking. This means that all users of Daum News are interested in similar issues, but what they read is different; less than 10% of news items were in common on average for those pairs of different sex groups while the average ρ is 0.65 for these pairs. We also find that two groups, Age10_Male and Age10_Female, are generally less alike with other groups, confirming the existence of an age gap between 10s and others. We also note that while the Age50_Male group has very different topical interests, it has similar issue preferences with other groups.

4.4 By Angle

We firstly observed that demographic groups show such different news consumption patterns from the low news item's similarity. Then, the high topic and issue similarity scores tell us that the overall news consumption is largely driven by current issues, but that still leads to groups having distinct news consumptions. This suggests that news consumptions even for one particular issue may be very different across groups consuming news articles with different angles. The "Angle" of a news story is similar to the concept of a follow-up story that gives readers new details followed by the background that is found in the initial story. This can include different event focus or other viewpoints. The news angle is different from framing but can be a combination of framing. We now examine the different news angles of one particular issue the groups are consuming.

For this analysis, we selected the most popular issue in our dataset, the Middle East Respiratory Syndrome (MERS) outbreak in South Korea. The outbreak lasted for a month and a half starting on 20th May, and from the outbreak, a total of 186 cases were infected, with a death toll of 36. Due to the outbreak, 2,208 schools were temporarily closed, and 16,693 people were quarantined.

The MERS outbreak was a deviant event, and we find all ten groups have at least one news item about MERS. However, the volume of news items about MERS is different across groups. News items about MERS are more popular in female groups than in male groups–on average, the female group has 427.4 popular news items about MERS while that of the male group is 123.8.

We then quantify the differences of MERS news consumption in terms of the content between two groups. To do this, we firstly select words that appeared at least 10 times across all news titles about MERS, which is 264 words out of 19,196 words. Then for each group, we rank those words based on the number of its occurrences in the popular news about MERS. We test the similarity in terms of news titles between two groups by computing the Spearman's ranking correlation coefficient. This will tell us which two groups have

Group	Distinctive words
Age10_Female	80s, Manpower, Everyone, Driven by, Increased, Died, Death, Self-quarantine, Getting on, Face the crossroads, Defenseless, Government, 19 people, Investigation, Virus, Heat wave, Still, 9 people
Age20_Female	Jeju island, Cases, Positive test result, Seoul, Possibility, Refuser, Wild ticks, Cytokine storm, Contact, Female, A patient, Virus, Trot, Condition
Age30_Female	Pregnant women, Occur, A patient, High fever, 2 people, Adding, Mask, Entrance, Gloves, Local hospitals, Close contact, Increased, 180 people, Cured, The number of patients, A public servant, Male
Age40_Female	Student, Infection, School, A patient, Son, On leaves, Elementary school students, Medical team, False charge, Grandmother, Visited, Guardian, Closed down, 100 places, Hospital, Kindergarten, Children, Teaching
Age50_Female	Samsung Seoul Hospital, Partially closed down, Infection, Gandong Sungsim Hospital, Large hospitals, Working, Epicenter, Concentration, 9 days, This week, Stable situation, Mystery, Diagnosis, Keep the principle, Remaining, Ambush, Close-packed, Go through, Jongdo Lim
Age10_Male	Believable, Diverse, A MERS map by a programmer, Imported cars, Fall down, Early next week, The executives, Hongik University, Confirmed infected patients, Removed, Jonlo, How far, For taking metro, Mockery flyers
Age20_Male	President, Won-soon Park (The mayor of Seoul), Jae-myeung Lee (The mayor of Sungnam), On leaves, Announcement, 35 times, WHO, SARS, Direct, Misreport, Doctors, Hyungpyo Moon, Entrance, Troll, Soldiers, Qatar, Stigma
Age30_Male	Won-soon Park, Jae-myeung Lee, Samsung Seoul hospital, Mu-sung Kim (Floor leader of ruling party), SARS, The mayor of Sungnam, President, WHO, Trot, Troll, Problem, Exterminator, Standard procedure, Despite of being a director, Our nation, 35 times
Age40_Male	President Park, False propaganda, A boy, Over-reaction, Firmly, Shepherd, Make, Should not do it, Rumor, Damage, Response, Provided, Separate, Many people, Marine Police, Step on, Last year, President, Disgust, Medical schools
Age50_Male	Replacement, Lacks proper response, Vice minister, Sorry, Lowering, Kyo-ahn Hwang (Prime minister), Delayed, Ruling and opposition parties, Shaking hands, Response, Briefing, 41 people, Presidential candidate, Debut, Site, Political issue, As scheduled

Table 1: Most discriminative words in news title about MERS outbreak across demographic groups, ranked by *phi*

the most similar set of popular news items about MERS. Figure 2(d) shows the results as a heatmap. All pairs of rankings are statistically significantly different ($p < 0.05$).

From the heatmap, we observe that 1) the popular news are similar within the same sex groups than within the different sex groups, 2) Female groups are more similar to each group than Male groups are; 3) age differences matter, except Age10_Male and Age10_Female groups. Interestingly, all three of these observations are also found in our previous analysis of comparing actual news items in Section 4.1. All these results lead us to conclude that all groups are generally interested in similar topics or issues; however, for the same issue, they are attracted to different news articles, leading to the big differences of popular news among groups.

To gain insights into how the popular news items about MERS differ between different demographic groups, we extract the most discriminative words in news titles for each group. We focus on the group-specific words of news titles, specifically on those with a high Chi-square score [2] for discriminating between one group and others (e.g., Age10_Female vs. Non-Age10_Female (all other groups)). Table 1 shows the top 20 words ranked by *phi*, the Chi-square test statistics. Two authors of this work translated Korean words to English words and removed a word if it overlaps with another. Some interesting differences were observed. Overall, Female groups are likely to check the status of the MERS outbreak, such as how many people are infected (the number of patients, death, this week), the symptoms of MERS (high fever, Cytokine storm), and the protection against MERS (mask, gloves). The Age30_Female group showed an interest in news about pregnant women who had been diagnosed with MERS and other women's cases. The Age40_Female group in particular was more interested in the status of closed schools and other education-related topics. On the other hand, the Male groups were more interested in the political issues surroungding the MERS outbreak, the accusation of the government in its response to the MERS outbreak (e.g., ruling and opposition parties, the lack of a proper response, misreporting, false propaganda), and the responsibility of politicians (President Park, the mayor of Seoul and Sungnam).

5. DISCUSSION AND CONCLUSION

To the best of our knowledge, this is the first study to conduct a multidimensional analysis of the news consumption of different demographic groups on a nationwide scale. News topic preferences based on age and sex exist among South Koreans. We went beyond topic analysis and examine issue and angle preferences, finding topic and issue preferences are similar across groups, but angle preferences are not. This means angle differences can explain the strikingly low numbers of common news items across groups, whereas the topic and issue differences cannot. In summary, groups are generally interested in similar issues, but they read news articles from different angles, indicating that angles make what news consumption look different.

Although our current study is limited to South Koreans due to the availability of detailed data, in principle, our results are applicable to any country if one translates Korean words into other languages. We are currently working on building a language model that predicts the preferences of different demographic groups. This will be especially helpful if the demographics of readers is not readily available on news sites.

6. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[3] W. Chu and S.-T. Park. Personalized recommendation on dynamic content using predictive bilinear models. In *Proceedings of the 18th international conference on World wide web*, pages 691–700. ACM, 2009.

[4] H. I. Chyi and A. M. Lee. Online news consumption: A structural model linking preference, use, and paying intent. *Digital Journalism*, 1(2):194–211, 2013.

[5] L. dâĂŹHaenens, N. Jankowski, and A. Heuvelman. News in online and print newspapers: Differences in reader consumption and recall. *New Media & Society*, 6(3):363–382, 2004.

[6] S. Knobloch-Westerwick and S. Alter. The gender news use divide: Americans' sex-typed selective exposure to online news topics. *Journal of Communication*, 57(4):739–758, 2007.

[7] J. Liu, P. Dolan, and E. R. Pedersen. Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 31–40. ACM, 2010.

[8] M. J. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5-6):393–408, 1999.

[9] Pew Research Center for the People and the Press. News audiences increasingly politicized-Online news audience larger, more diverse: Biennial media consumption 2004. http://www.people-press.org/2004/06/08/news-audiences-increasingly-politicized/, 2004. Retrieved 22nd July 2016.

[10] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 2012.