Multidimensional Analysis of the News Consumption of Different Demographic Groups on a Nationwide Scale

Jisun An(✉) and Haewoon Kwak

Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
{jisun.an,haewoon}@acm.org

Abstract. Examining 103,133 news articles that are the most popular for different demographic groups in Daum News (the second most popular news portal in South Korea) during the whole year of 2015, we provided multi-level analyses of gender and age differences in news consumption. We measured such differences in four different levels: (1) by actual news items, (2) by section, (3) by topic, and (4) by subtopic. We characterized the news items at the four levels by using the computational techniques, which are topic modeling and the vector representation of words and news items. We found that differences in news reading behavior across different demographic groups are the most noticeable in subtopic level but neither section nor topic levels.

Keywords: News consumption · Online news · News media · News topic · Daum · News portal · Demographics · Gender differences · Age differences

1 Introduction

Demographics play an important role in news consumption. What women in their fifties read is very different from that of men in their twenties. Understanding such differences in news consumption can potentially help journalists to pitch news articles better, help editors to decide which ones to put on the front page, and help computer scientists design new algorithms for recommending articles. That is the reason why news consumption of demographic groups has been actively studied in both the domains of journalism study and computer science [14,18,24,31,39,40].

The previous literature regarding news consumption in the study of journalism has mainly focused on the gender differences in the consumption of news genres [18,24,40] primarily due to the lack of other detailed data. On the other hand, in computer science, previous studies have mainly focused on developing models for predicting clicks and for news recommendations at an individual level with large-scale data [14,31,39].

G.L. Ciampaglia et al. (Eds.): SocInfo 2017, Part I, LNCS 10539, pp. 124–142, 2017.
DOI: 10.1007/978-3-319-67217-5_9

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

For this study, we collected and analyzed the daily top 30 news items for each gender (male and female) and age group (10s, 20s, 30s, 40s, and 50s) in Daum News, the second most popular news portal service in South Korea, for the entire year of 2015. The number of the unique news items collected is 103,133. Daum News can have the accurate, not self-reported, information on the user's age and gender based on one's social security number. This practice is not common in the Western web services. In South Korea, to join a website, it is mandatory for identity verification to provide the social security number that contains your birth year and gender. Also, Daum News has a strong user base that reads news with a logged-in status mainly because Daum News offers a wide range of services, which include e-mail, Internet community, or messenger, for example, based on the logged-in status.

2 Related Work

2.1 Sex-Typed Media Preference

Sex-typed media preference has long been investigated for various types of media. Regarding movie and TV genres, studies have found that women are more likely to watch tragedies, soaps, dramas, medical serials, and romances; men, on the other hand, tend to prefer horror, sports, and action and adventure content [21,37,38,41,42]. Psychological research suggests that such sex-typed media selections might well be rooted in societal gender stereotypes – men are expected to achieve more, and women are expected to interact more [16].

Scholars of the study of journalism have examined how demographics, such as sex or age, relate to news-seeking behavior and news preferences [15]. When attending to news, the sexes typically pursue remarkably different interests – in terms of topics, men tend to follow news on politics, sports, and business and finance, whereas women turn to news about community and health issues [18,40]. Also, women read more about social/interpersonal issues than men, and men read more about achievement/performance than women [24]. Scholars suggested that the origin of gender difference in news consumption is not considered from biological differences but from the psychological traits led by the sex-typed socialization [33].

While news preferences have been extensively studied for the news section, little is known about the topic or subtopic preferences of different demographic groups' news consumption. The key reason for this oversight is mainly due to a lack of data. Most of the previous work is based on surveys or experiments in

the laboratory setting [18,24,40]. By contrast, our work relies on the longitudinal data collection where a huge number of Korean Internet news readers are unobtrusively monitored.

2.2 News Related Research with Large-Scale Data

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 addressed news-related questions but with different interests and approaches from scholars in journalism or communication studies. News-related research by computer scientists has predominantly focused on modeling news sharing behavior [3–5], news diffusion [1,7,17,19,26,29,36,44,49], and modeling the relative prominence of items or topics [2,6,8,13,25,28].

Computer scientists also have exploited news consumption patterns of individuals mainly for building a better news recommendation system to give readers a personalized experience when reading the news. Systems that make recommendations according to demographic classes were initially introduced [39]. More recently, the demographic information has often been used as a feature of those models [14,31]. However, the demographic differences in news consumption have not been fully uncovered, particularly in different dimensions of news such as topic or subtopic.

3 Data

3.1 News Consumption in South Korea

Online news consumption in South Korea has increased drastically. About 86% of South Korean people access news online at least once a week[1]. Given the 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 South Korea have been eager to publish their content via portals for years. In 2015, Naver and Daum 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 reasonably say that news providers and the news readers of Daum News are representative of the general South Korean news media and population, respectively.

3.2 Data Collection

Daum is the second largest web portal in South Korea, followed by Naver. Daum plays a significant role in providing a place for accessing online news to South Korea; 41% of South Koreans (24.6 M users) access Daum news on a weekly basis. Daum News provides different ways to explore news articles, for example, by its

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

recency, by current issues, by regions, or by popularity 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 note that Naver does not provide a ranked list of news articles by different demographic groups, and thus, we focus on Daum News data even though it is the second largest news portal in South Korea.

We crawled the top 30 most popular articles for different age and gender groups of each day for a one-year period (01/01/2015−31/12/2015). We carefully designed our crawlers not to degrade the performance of Daum's web servers. 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 item 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.

4 Methodology

We need to determine the section, topic, and subtopic of the news items we collected. Next, we will briefly describe the methodology that we adopt for the analysis.

4.1 Four Dimensions of News

Defining the operational coding scheme for news has been a central issue for communication studies. In this study, we introduce four dimensions of news (sections, topics, subtopics, and individual news items) by which we investigate news consumption patterns. The diagram of the four dimensions of news is illustrated in Fig. 1. The first dimension is the news section (e.g., society, politics, etc.), which is a category of news often adopted in newsrooms and which has been popularly used for sex-typed news preference studies. Next, a news topic refers to a specific happening (e.g., a MERS outbreak in South Korea or child abuse case in a daycare center). Within a topic, we further explore subtopics by distinguishing different aspects of a topic. Finally, there is a dimension where

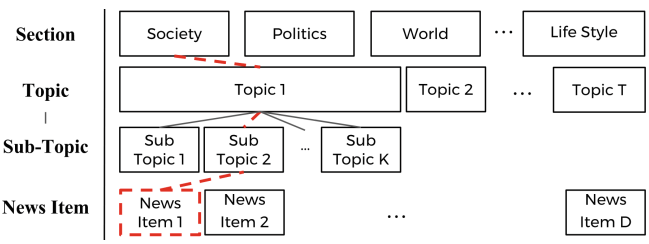


Fig. 1. Four dimensions of news.

all individual news items are aligned. We then can characterize the entire news collection (e.g., a news collection of South Korea in 2015) according to the four dimensions.

4.2 Categorizing News: Section

The most common way of categorizing news is perhaps to use a news section as defined by the news media. Especially for online news, which news section a news article belongs to can be inferred from 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 the collected URLs and extracted the section information.

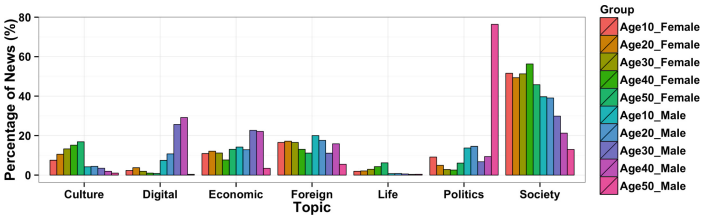


Fig. 2. Group distribution by section.

Figure 2 shows the proportion of news items for each demographic group for each news section. In our dataset, Daum News has seven different news sections, which are Culture, Digital, Economic, Foreign, Life, Politics, and Society. We note that the Entertainment and Sports sections are not included in the dataset. We observe that female groups are more alike than male groups in terms of news section preference. For female groups, Society and Culture were the two most popular news sections. For Age10_Male and Age20_Male groups, Society, Foreign, and Politics were more popular than the others. Then, for the Age30_Male and Age40_Male groups, their interests were Digital (i.e., technology news), Economics, and Politics. We also find one noticeable behavior of the Age50_Male group: 78% of their top 30 most popular news items over a year are about Politics. These news section preferences align well with previous work on sex-typed news consumption observed internationally [18,40].

4.3 Categorizing News: Topic

The news section category provided by Daum News abstractly captures the high-level topic of a news item. In this study, we go beyond a mere news section preference and examine whether the gender- or age-specific topic or subtopic 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 under Society, but they are two different *topics*.

We automatically discover a topic-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) [23], Latent Dirichlet Allocation (LDA) [11], and Hierarchical Dirichlet Processes (HDP) [46] can be employed to induce topics from the set of news titles. We manually compare the three methods in terms of the interpretability of the induced topics and the quality of the clustered news titles for each of the included topics. For two slices of subsets, May and October 2015, we examined the top 20 topics resulting from the three methods with their top 30 words. For our dataset, pLSI's top topics were mapped more clearly with news events in the corresponding month than LDA's and HDP's topics were. One possible reason is that LDA and HDP were penalized more for modeling topics with short titles than pLSI was. Thus, we use pLSI to detect candidate topics. There are also specialized topic models for short texts, such as a biterm topic model (BTM) [48], and we leave it as a future work to improve our topic categorization method.

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 topic even though one event 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 build a pLSI model for each of the monthly datasets. Each pLSI model induces 100 topics, giving us 1,200 candidate topics in total. Each topic is represented by word distributions. We then aggregate similar candidate topics by examining the representative keywords for topics (words with the highest probabilities). In our case, we aggregate topics if they share more than three top words. After this, we classify each news title as one of these candidate topics based on the score given by the pLSI model. Finally, we split every candidate topic further into multiple topics by considering the publication times of news items. Only news items published for days in a row are tagged as the same topic, resulting in 2,122 topics. News titles without any matching candidate topic are considered stand-alone topics. Altogether, we have 41,452 topics.

Across all topics, the most popular news event is the Middle East Respiratory Syndrome (MERS) outbreak in South Korea. We find 2,756 matching news items regarding MERS. The most frequent words are MERS, a vice prime minister, confirmed patients, infected, hospital, Daejeon, tourists, etc.

Validation of Topic Categorization. For the purpose of validation, we cross-matched the set of news items extracted by the topic-model-based clustering method with the set of news items extracted by a keyword-based method. We focus on one particular news topic, MERS outbreak in South Korea. The outbreak lasted for a month and a half starting on May 20th. A total of 186 cases occurred during the outbreak, with a death toll of 36. Due to the outbreak, 2,208 schools were temporarily closed, and 16,693 people were quarantined. MERS was the most sensational news event in South Korea in 2015. We extract news items

whose title contained the word "MERS", which results in 961 news items. By including more keywords, we may be able to extract more news items. However, we use a single word to be sure that all the retrieved news items are relevant to the news event. We find that 97% of news items about the MERS outbreak (929 out of 961 news items) overlap with those by the topic modeling based method. Then, to examine the relevance of those articles without the word "MERS," we randomly select 100 articles that do not explicitly include the word "MERS" and examine what they are about. We find that all are relevant to the MERS. This indicates that our method can extract news items even when the title of news does not include a key topic word. However, a set of topic words (not one word) will also be able to retrieve all relevant news items.

4.4 Categorizing News: Subtopic

Once we have a set of news articles on a certain topic, we further need to group them by subtopic. To this end, we train our data to represent each news item on a vector space and then cluster news items. Semantic vector space models of language represent each word with a real-valued vector. Firstly introduced by Hinton [22], the methods have been extended using neural network [9,45] and applied for the practical uses. In recent years, Milkolvo's skip-gram and distributed bag-of-word (DBOW) models [34] are popularly used for learning vector representation of words and documents due to its computational efficiency. In the skip-gram model, the objective is to predict a word's context given the word itself, whereas the objective in the DBOW model is to predict a word given its context.

More recently, the concept of embeddings has been extended beyond words to a number of text segments, including phrases [35], sentences and paragraphs [27], and documents. Adopting the document representation method, we learn distributed representations of news items in our Daum news collection. Each of news items is represented as low-dimensional vectors and are jointly learned with distributed vector representations of words using a DBOW model explained in [27]. In this vector space, two news items of semantically similar meaning are located nearby.

In our news embeddings (henceforth *News2Vec*), every news item is mapped to a unique vector in a matrix that represents news items, and every word is mapped to a unique vector in a matrix that represents words. We denote by $\mathbf{N2V}$ the $s \times f$ matrix of s which is the sum of n news items (N_1, N_2, ..., N_n) and f dimensions (F_1, F_2, ..., F_f). A great advantage of learning distributed representation vectors for news in this way is that the algorithm is not sensitive to news item length and does not require specific tuning for word weights. The row of the matrix $\mathbf{N2V}$, $\mathbf{N_i}$, is a vector of f dimensions representing the i-th news item. The dimension of vector f is set to 200, and the model is trained with 40 epochs.

Once we have News2Vec, we apply the hierarchical clustering method using Ward's method [43] to this resulting matrix to cluster news items of the same

subtopic. The hierarchical clustering method builds a dendrogram among entities. Then, one can cluster entities based on the dendrogram. Its main advantage is that the dendrogram is computed only once regardless of the number of clusters of interest. Once the dendrogram is built, we can simply choose the number of clusters (k) of our interest.

Validation for Subtopic Categorization. To evaluate the News2Vec based subtopic clustering method, we prepared a corpus in which each news item is labeled by two authors. We used 961 news items about MERS extracted by the keyword-matching method. Then, we manually classify news items by subtopics. We conduct a qualitative content analysis to develop a taxonomy of subtopics for MERS news.

Following an open-coding method [47], we identify the subtopics of MERS news in a two-phase process. We first read titles and descriptions of 100 news items to develop an initial coding scheme and then used an affinity diagramming technique [10] to iteratively develop a classification scheme for subtopics until a new subtopic did not emerge. Table 1 lists the resulting 10 subtopics. The individual authors manually classified all news items into one of the subtopic categories. Using the Delphi method [30], after each researcher independently coded the titles, we then iteratively compared and recoded the news items as necessary until we came to an agreement.

The ten subtopics regarding the MERS with the number of corresponding news items are listed in Table 1. ST1 "Reporting new cases" was the most popular subtopic with 254 news articles in our data set, followed by ST7 "Responsibility of Government" with 153 news items and ST9 "Economical Consequences" with 106 news items.

We then use this labeled data for the evaluation of our subtopic categorization method. We first learn News2Vec using entire Daum News data. Then, we cluster

Table 1. Ten subtopics regarding the MERS outbreak in South Korea in 2015. The total number of news items is 961.

	Subtopic	# Items
ST1	Reporting new cases	254
ST2	Track the path of outbreak	67
ST3	Schools closing down	34
ST4	Factual information	54
ST5	Verification of rumors	24
ST6	Problem of medical systems	71
ST7	Responsibility of government	153
ST8	Government' solution	103
ST9	Economical consequences	106
ST10	Societal problems	95

those 961 news items into ten groups ($k = 10$) using vectors from the resulting News2Vec. We evaluate the resulting subsets of news items with manually tagged clusters. For each detected subset, we find the best matched manual cluster based on the proportion of matching news items. Across ten subsets, our News2Vec subtopic classification achieved an 82.2% matching rate on average where the maximum matching rate is 92.8% and the minimum is 77.2%.

In our method, selecting the k is challenging. Here, we propose one possible solution to assist in the k selection procedure. The idea is that the average similarity scores for all pairs of news items within the same subtopic (S_{within}) should be smaller than across the subtopics (S_{across}). When news items are in vector representation, one can use any distance measures, such as Euclidean distance or cosine similarity, to measure the similarity score between two news items. Those two values, once found empirically, can play a role as thresholds for selecting the k for different topics. In our evaluation data set, S_{within} is 0.02 and S_{across} is 0.15 when using cosine similarity.

For further evaluation, we use these two values to find the subtopics of another topic, "Daycare child abuse." For 342 news items regarding the topic, we find k equals five. The manual inspection reveals the following five subtopics emerged: (1) What the teacher did to a child; (2) How cruel the teacher is; (3) Investigation and prosecution; (4) A new regulation on CCTV installation at daycare centers; and (5) Other cases of child abuse.

5 Group Differences in News Consumption

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

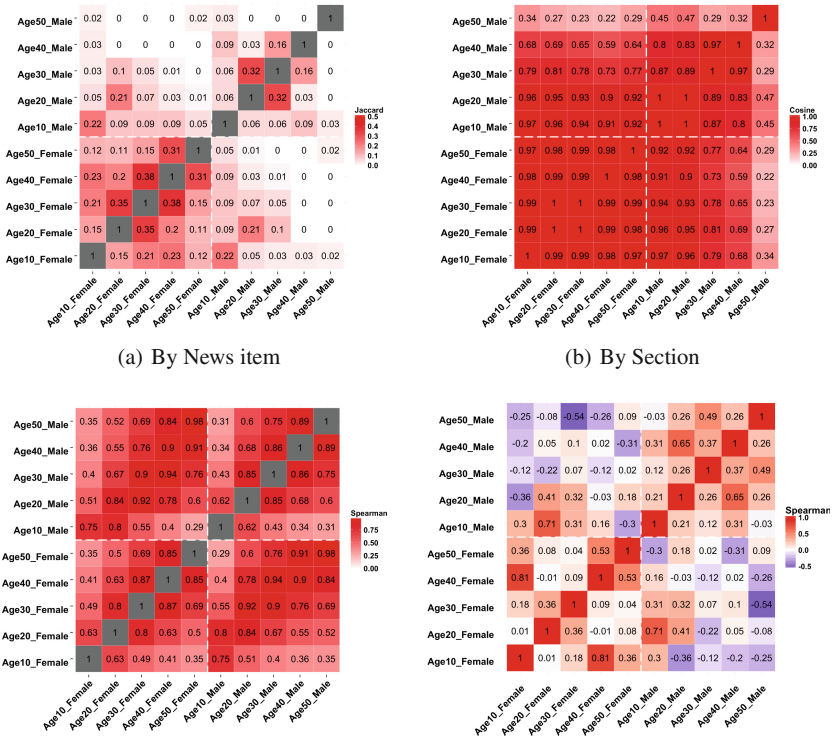
5.1 By News Item

As a first attempt to compare the news consumption of different groups, we look at actual news items. We measure the similarity among groups based on commonly consumed news items (by their unique news IDs) among the top 30 articles consumed by each group over a one-year period. We use Jaccard Similarity to compute group similarity. For the two sets, A and B, the Jaccard Similarity is given by: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. In 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 3(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 the Age30 or Age40 than the Age20 same-sex groups. In Fig. 2, 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.



(a) By News item (b) By Section

(c) By Topic (d) By Subtopic

Fig. 3. [Zoom-able in PDF] Heatmap showing the similarity across groups based on comparison of (a) news items – Jaccard similarity scores, (b) sections – Cosine similarity between section vectors, (c) topics – Spearman's rank correlation test of two lists of news topics ranked by their lifespans, and (d) subtopics – Spearman's rank correlation test of two lists of subtopics ranked by their number of news articles about the MERS outbreak.

5.2 By Section

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 sectional 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 vector $S_g = (w_{1,g}, w_{2,g}, \ldots, w_{s,g})$ in which each dimension corresponded to a predefined section from Daum News where $s = 7$ in our case (see 7 sections in Fig. 2). The weight $w_{s,g}$ was computed by the proportion of the news items in the section s for the group g. We then computed the cosine similarity between two vectors to compare the sectional interests of two groups. The results are shown in Fig. 3(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 section is similar in each group. A similar pattern is also observed for male groups but to a lesser extent, and with one exception, Age50_Male, which shows a very different sectional 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 Fig. 2).

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 sectional interests distinct from those of female groups. This partly supports the traditional sex-typed news consumption theory – our data set also shows different sectional preferences in different gender groups. However, we find such differences are driven more by Age30_Male and Age40_Male and less by Age10_Male and Age20_Male.

In summary, the sectional 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 fact that groups largely share sectional interests is still surprising. We now move onto the similarities in the topics that different groups consume.

5.3 By Topic

Given that sectional 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 topics – older people might read more about “Baby killer” while young people read more about “Violence at school.”

To investigate such topic-specific differences in news consumption, we map each news item to a specific topic. The topics are identified by the method we described in Sect. 4.3. Then, we quantify the importance or the level of attention to a specific topic for a group by computing the lifespan. We define the lifespan of a topic as the longest period of time when that topic appeared on the top 30 list for each group.

We then measure the similarities between groups based on the importance of the different topics. We select topics that are consumed by at least two groups, resulting in 36,134 topics and compute each topic's lifespan for each group. This gives us a ranked list of topics for each group, and we use Spearman's rank correlation coefficient (ρ) to compare the two ranked lists. Figure 3(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 that age differences play important roles in news consumption – a similar pattern was also found when looking at common news items in Sect. 5.1. Given that a pair of different sex groups have few common news items consumed, the high similarity between two ranked list of topics ($\rho > 0.8$) is striking. This means that all users of Daum News are interested in similar topics, but what they read is different; less than 10% of news items on average were in common 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 more different from other groups, confirming the existence of an age gap between 10-year-olds and others. We also note that while the Age50_Male group has very different sectional interests, it has similar topic preferences to those of other groups.

5.4 By Subtopic

We firstly observed that demographic groups show such different news consumption patterns at news item level. Then, the high similarity scores at section and topic levels tell us that the overall news consumption is largely driven by current issues. However, groups still have distinct news consumption patterns. This suggests that news consumption even for one particular topic may be very different across groups.

For this analysis, we use our evaluation data set and focus on the subtopic consumption regarding the MERS outbreak. The MERS outbreak was a deviant event, and all ten demographic 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 in MERS news consumption in terms of the content between two groups. To do this, for each group, we rank the subtopics of MERS outbreak in Table 1 by the number of news items. Then, we test the similarity between two groups by computing the Spearman's ranking correlation coefficient. This will tell us which two groups have the most similar consumption

of subtopics about MERS. Figure 3(d) shows the results as a heatmap. All pairs of rankings are statistically significantly different ($p < 0.05$).

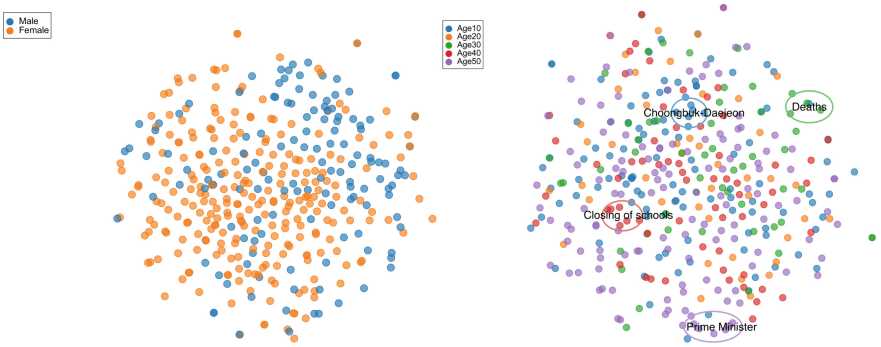
From the heatmap, we observe that (1) the popular news is more similar within the same sex groups than within the different sex groups; (2) female groups are more similar to each group than male groups are; and (3) age differences matter, except in the Age10_Male and Age10_Female groups. Interestingly, all three of these observations are also found in our previous analysis that compares actual news items in Sect. 5.1. All these results lead us to conclude that all groups are generally interested in similar news sections or topics; however, for the same topic, they are attracted to different subtopics, leading to the big differences in popular news among groups.

To gain insights into how 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 *phi* score, the Chi-square test statistics [12], for discriminating between one group and others (e.g., Age10_Female vs. Non-Age10_Female (all other groups)). Table 2 shows the top 20 words ranked by *phi*. Two authors of this work translated Korean words to English words. 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 surrounding the MERS outbreak, the accusations towards the government's 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.

News2Vec offers an opportunity to visualize each news item in the vector space by applying t-SNE, a widely-used dimensionality reduction based on manifold learning [32]. Figure 4 shows where each news item (colored circle) consumed by each gender (Fig. 4(a)) or age (Fig. 4(b)) group is located in the vector space. In the figure, closed circles are that news items have similar representations in the vector space and thus fall in similar subtopics. For the clarity of the visualization, we focus on the news items consumed by a single demographic gender or age group. Figure 4(a) shows a better clustered structure than Fig. 4(b), meaning that gender difference is well aligned with the difference of vector representation of news items than age difference. In Fig. 4(b), we can also see some clustered structures, such as groups of green circles (news on patients' deaths read by 30s), blue circles (news on Choongbuk-Daejeon regions read by 10s), red circles (news on the closing of schools read by 40s), and purple circles (news on the prime minister read by 50s) from the top to the bottom, while colored circles

Table 2. Most discriminative words in news title about MERS outbreak across demographic groups, ranked by *phi*

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, Samsung Seoul hospital, Mu-sung Kim (Floor leader of ruling party), SARS, The mayor of Sungnam, President, WHO, Trot, Troll, Problem, Exterminator, Standard procedure, Responsibility, 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, Political issue



(a) Gender Difference (b) Age Difference

Fig. 4. t-SNE visualization of news items colored by gender (a) and age (b) (Color figure online)

are mostly dispersed over the space mostly. This is another evidence that gender difference is more noticeable than age difference at the subtopic level.

6 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. Differences in news consumption between different demographic groups exist among South Koreans. We look into news consumption at different levels and find that section and topic preferences are similar across groups, but subtopic preferences are not. This means that only the behavioral differences in the subtopic level can explain the strikingly low numbers of common news items across different demographic groups, whereas the differences in the news section or topic levels cannot. In summary, while different demographic groups are interested in similar topics, they read news articles belonging to different subtopics, indicating that subtopics make the news consumption of the different groups different. For the following studies, our work suggests that the differences between demographic groups should be examined at the appropriate level.

One potential limitation is that we analyze a single news portal service, even though it is an extremely popular service in South Korea. To the best of our knowledge, Daum is the only data source where (1) the user-registered demographic information is credible, and (2) the user base spans all generations and parties. We are willing to extend our approach to new data sources that satisfy above two conditions so that we can find demographic differences in news consumption.

We note that all the users are exposed to the same layout and the same items if they visit the website at the same time. Thus, the ways to show news, such

as news clustering or adaptive layout, cannot selectively influence on a certain user segment.

While the differences in news consumption between demographic groups can be explained partly by different interests on subtopics, there could be other latent factors such as sentiments or frames of news. For instance, Grabe and Kamhawi found that men recognize and respond more to negatively framed messages, while women are more aroused by and engaged with positively framed messages [20]. By adding another dimension to our news dimensions, we can understand human behavior on consuming news in depth. In this work, we examined the most critical aspects of news, which are section, topic, and subtopic, and saved the other dimensions for the future work.

Our results also bring the practical implications for news organizations. Our characterization of user behavior allows them (1) to gain a better understanding of what people consume and (2) to produce more relevant and engaging content for different demographic groups. Such demographic-based profiling can help tackle the cold-start problem for new users. By simply knowing the gender and age of a new user, one can provide a better user experience when reading online news. The opposite direction of the inference can also be useful. Given the first few sets of news articles, one can infer the demographics of the reader. This will be particularly useful if the demographics of users are not readily available on news sites.

References

1. Adar, E., Adamic, L.A.: Tracking information epidemics in blogspace. In: Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 207–214. IEEE Computer Society, Washington, D.C. (2005)
2. Ahmed, M., Spagna, S., Huici, F., Niccolini, S.: A peek into the future: predicting the evolution of popularity in user generated content. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 607–616. ACM (2013)
3. An, J., Cha, M., Gummadi, K., Crowcroft, J.: Media landscape in Twitter: a world of new conventions and political diversity. In: ICWSM 2011: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 18–25 (2011)
4. An, J., Quercia, D., Cha, M., Gummadi, K., Crowcroft, J.: Sharing political news: the balancing act of intimacy and socialization in selective exposure. EPJ Data Sci. **3**(1), 12 (2014)
5. An, J., Quercia, D., Crowcroft, J.: Partisan sharing: Facebook evidence and societal consequences. In: CoSN 2014: Proceedings of the 2nd ACM Conference on Online Social Networks, pp. 13–24. ACM (2014)
6. Asur, S., Huberman, B.A., Szabo, G., Wang, C.: Trends in social media: persistence and decay. In: ICWSM 2011: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 434–437. Association for the Advancement of Artificial Intelligence (AAAI), Menlo Park (2011)
7. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.A.: The role of social networks in information diffusion. In: WWW 2012: Proceedings of the 21st International Conference on World Wide Web. ACM, New York (2012)

8. Bandari, R., Asur, S., Huberman, B.A.: The pulse of news in social media: forecasting popularity. In: ICWSM 2012: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, pp. 26–33. Association for the Advancement of Artificial Intelligence (AAAI), Menlo Park (2012)
9. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
10. Beyer, H., Holtzblatt, K.: Contextual Design: Defining Customer-Centered Systems. Elsevier, Amsterdam (1997)
11. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
12. Casella, G., Berger, R.L.: Statistical Inference, vol. 2. Duxbury, Pacific Grove (2002)
13. Cheng, J., Adamic, L.A., Dow, P.A., Kleinberg, J., Leskovec, J.: Can cascades be predicted? In: WWW 2014: Proceedings of the 23th international conference on World Wide Web, pp. 925–936. ACM, New York (2014)
14. Chu, W., Park, S.T.: Personalized recommendation on dynamic content using predictive bilinear models. In: WWW 2009: Proceedings of the 18th International Conference on World Wide Web, pp. 691–700. ACM (2009)
15. Chyi, H.I., Lee, A.M.: Online news consumption: a structural model linking preference, use, and paying intent. Digit. Journal. **1**(2), 194–211 (2013)
16. Cross, S.E., Madson, L.: Models of the self: self-construals and gender. Psychol. Bull. **122**(1), 5 (1997)
17. Dow, P.A., Adamic, L.A., Friggeri, A.: The anatomy of large Facebook cascades. In: ICWSM 2013: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, pp. 145–154. Association for the Advancement of Artificial Intelligence (AAAI), Menlo Park (2013)
18. d'Haenens, L., Jankowski, N., Heuvelman, A.: News in online and print newspapers: differences in reader consumption and recall. New Media Soc. **6**(3), 363–382 (2004)
19. Goel, S., Anderson, A., Hofman, J., Watts, D.J.: The structural virality of online diffusion. Manag. Sci. **62**(1), 180–196 (2016)
20. Grabe, M.E., Kamhawi, R.: Hard wired for negative news? Gender differences in processing broadcast news. Commun. Res. **33**(5), 346–369 (2006)
21. Hansen, C.H., Hansen, R.D.: Music and music videos. In: Media Entertainment: The Psychology of its Appeal, pp. 175–196 (2000)
22. Hinton, G.E., McClelland, J.L., Rumelhart, D.E.: Distributed representations. In: Rumelhart, D.E., McClelland, J.L., PDP Research Group CORPORATE (eds.) Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 77–109. MIT Press, Cambridge (1986). ISBN 0-262-68053-X. http://dl.acm.org/citation.cfm?id=104279.104287
23. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM (1999)
24. Knobloch-Westerwick, S., Alter, S.: The gender news use divide: Americans' sex-typed selective exposure to online news topics. J. Commun. **57**(4), 739–758 (2007)
25. Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., Kustarev, A.: Prediction of retweet cascade size over time. In: CIKM 2012: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2335–2338. ACM, New York (2012)
26. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: WWW 2010: Proceedings of the 19th International Conference on the World Wide Web, pp. 591–600. ACM, New York (2010)

27. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: ICML 2014: Proceedings of the 31st International Conference on Machine Learning, pp. 1188–1196. PMLR (2014)
28. Lerman, K., Hogg, T.: Using a model of social dynamics to predict popularity of news. In: WWW 2010: Proceedings of the 19th International Conference on World Wide Web, pp. 621–630. ACM (2010)
29. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: KDD 2009: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 497–506. ACM, New York (2009)
30. Linstone, H.A., Turoff, M., et al.: The Delphi Method: Techniques and Applications, vol. 29. Addison-Wesley, Reading (1975)
31. Liu, J., Dolan, P., Pedersen, E.R.: Personalized news recommendation based on click behavior. In: IUI 2010: Proceedings of the 15th International Conference on Intelligent User Interfaces, pp. 31–40. ACM (2010)
32. Maaten, L.V.D., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**(Nov), 2579–2605 (2008)
33. Maccoby, E.E.: Gender as a social category. Dev. Psychol. **24**(6), 755 (1988)
34. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
35. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
36. Myers, S., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. In: Yang, Q., Agarwal, D., Pei, J. (eds.) KDD 2012: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 33–41. ACM, New York (2012)
37. Oliver, M.B.: Exploring the paradox of the enjoyment of sad films. Hum. Commun. Res. **19**(3), 315–342 (1993)
38. Oliver, M.B.: The respondent gender gap. In: Media Entertainment: The Psychology of its Appeal, pp. 215–234 (2000)
39. Pazzani, M.J.: A framework for collaborative, content-based and demographic filtering. Artif. Intell. Rev. **13**(5–6), 393–408 (1999)
40. Pew Research Center for the People, the Press: News audiences increasingly politicized-Online news audience larger, more diverse: Biennial media consumption 2004 (2004). http://www.people-press.org/2004/06/08/news-audiences-increasingly-politicized/. Accessed 20 July 2017
41. Potts, R., Dedmon, A., Halford, J.: Sensation seeking, television viewing motives, and home television viewing patterns. Pers. Individ. Differ. **21**(6), 1081–1084 (1996)
42. Preston, J.M., Clair, S.A.: Selective viewing: cognition, personality and television genres. Br. J. Soc. Psychol. **33**(3), 273–288 (1994)
43. Rokach, L., Maimon, O.: Clustering Methods. Springer, Boston (2005)
44. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: WWW 2011: Proceedings of the 20th International Conference on World Wide Web, pp. 695–704. ACM, New York (2011)
45. Schwenk, H.: Continuous space language models. Comput. Speech Lang. **21**(3), 492–518 (2007)

46. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Sharing clusters among related groups: hierarchical dirichlet processes. In: Advances in Neural Information Processing Systems, pp. 1385–1392 (2005)
47. Tesch, R.: Qualitative Research: Analysis Types and Software Tools. Psychology Press, Hove (1990)
48. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: WWW 2013: Proceedings of the 22nd International Conference on World Wide Web, pp. 1445–1456. ACM (2013)
49. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: ICDM 2010: Proceedings of the 2010 IEEE International Conference on Data Mining, pp. 599–608. IEEE Computer Society, Washington, D.C. (2010)