

텍스트 마이닝을 활용한 세대별 키워드 빅데이터 분석: 네이트판 10대 · 20대 · 30대 게시판을 중심으로

정 백[○], 배성원^{*}, 황보유정^{**}

[○]경희대학교 빅데이터응용학과,

^{*}경희대학교 빅데이터응용학과,

^{**}경희대학교 빅데이터연구센터

e-mail: {ylbaek[○], sl, hwangbo^{**}}@khu.ac.kr

Bigdata Analysis on Keyword by Generations through Text Mining: Focused on Board of Nate Pann in 10s, 20s, 30s

Baek Jeong[○], Sungwon Bae^{*}, Yujeong Hwangbo^{**}

[○]Department of Big Data Analytics, Kyung Hee University,

^{*}Department of Big Data Analytics, Kyung Hee University,

^{**}Big Data Research Center, Kyung Hee University

● 요약 ●

본 논문에서는 텍스트 마이닝 기법을 이용하여 MZ 세대를 이해하는 키워드를 도출하고자 한다. MZ 세대의 비중이 높아지면서, MZ 세대를 분석하려고 하는 많은 연구들이 수행되고 있다. 이에 본 연구에서는 MZ 세대를 이해하기 위하여 네이트 판의 연령별 게시판 크롤링을 통해 빅데이터를 수집하였다. 그리고 텍스트 마이닝 기법을 활용하여 10대, 20대, 30대의 각각의 키워드를 도출할 수 있었다. 본 논문에서 도출된 키워드는 이는 MZ 세대를 이해하는데 중요한 키워드로 볼 수 있을 것이다. 향후 연구로는 MZ 세대와 기성 세대를 비교하기 위하여 추가 크롤링을 통해 세대 간 비교 연구를 수행하고자 한다.

키워드: 텍스트 마이닝(text mining), 세대(generation), 빅데이터(bigdata), 네이트 판(nate pann)

I. Introduction

최근 사회적 이슈로 MZ 세대와 기성 세대간의 갈등이 심화되고 있다. 이러한 갈등은 지난 3월에 있었던, 20대 대통령 선거에서 생별이나 세대간의 갈등이 주요한 쟁점이 될 정도로 사회적으로 큰 이슈가 되고 있다. MZ 세대는 밀레니얼 세대와 Z를 합쳐서 부르는 말로, M은 밀레니얼 세대(Millennial Generation)로 1980년대 초 ~ 2000년대 초에 태어난 세대, Z는 Z 세대(Generation Z)로 1995년 이후에 태어난 세대를 의미한다[1]. 2022년 통계개발원에서 발표한 「KOSTAT 통계플러스」에 따르면 MZ 세대는 약 16,300천 명으로 국내 총 인구의 32.5%를 차지하는 것으로 나타났다. 이것은 MZ 세대의 수의 비중이 높다는 것을 의미하며, 머지않아 MZ 세대가 사회의 주류가 된다는 것을 시사한다. 이렇게 MZ 세대의 비중이 높아지고 있는 가운데, 많은 연구들은 MZ 세대를 분석하여, 새로운 서비스를 개발하거나, 세대간 갈등을 해결하는데 해결책을 찾고 있다 [2,3,4]. 본 연구에서는 MZ 세대의 특징을 분석하기 위하여, MZ 세대들이 많이 사용하고 있는 네이트 판을 크롤링하여, 연령대 별 게시판을 글을 수집하였다. 수집된 텍스트들을 텍스트 마이닝 기법을

사용하여 주요 키워드를 도출한 후, 10대, 20대, 30대 별 핵심 키워드를 비교하여 현재 MZ 세대들의 세대별 키워드를 도출하였다. 이를 통해 MZ 세대들을 이해하고 세대간 이슈를 해결하는데 나아갈 방향을 제시하고자 한다.

II. Method

세대별 키워드 분석을 위해 웹크롤링 기법을 활용하여, 네이트판 세대별 게시판(10대 이야기, 20대 이야기, 30대 이야기)의 최근 게시물을 각 게시판별 3,000개씩 수집하였다. 분석에는 R studio를 활용하여 먼저 형태소 분리를 통해 게시글을 명사 단위로 추출하였다. 가장 많은 키워드를 확인할 수 있는 빈도 분석과 이를 시각화한 워드클라우드 기법을 사용하였으며, 핵심 키워드 간의 사회 연결망의 확인을 위하여 연결중심성 분석과 위세중심성을 지표로 키워드간 네트워크를 확인하였다.

Table 1. 세대별 키워드 빈도(상위 30개) 및 TF-IDF

10대				20대				30대			
순위	keyword	n	TF-IDF	순위	keyword	n	TF-IDF	순위	keyword	n	TF-IDF
1	친구	379	69.83	1	사람	1535	69.78	1	사람	1638	47.47
2	사람	309	84.74	2	친구	1211	61.12	2	생각	1009	29.41
3	공부	164	48.06	3	생각	904	35.60	3	친구	823	37.46
4	엄마	155	33.01	4	남친	649	50.76	4	남자	758	51.41
5	학교	142	37.07	5	엄마	590	24.29	5	결혼	688	35.91
6	생각	141	42.51	6	남자	546	50.37	6	여자	683	43.45
7	남자	126	53.06	7	여자	499	45.45	7	얘기	631	15.87
8	남친	126	44.58	8	얘기	485	21.24	8	여친	481	24.93
9	인스타	110	44.40	9	연락	422	22.15	9	마음	446	20.45
10	얼굴	105	44.77	10	마음	330	19.31	10	30대	404	32.54
11	짹남	103	35.88	11	부모님	315	22.02	11	문제	397	14.15
12	시험	98	39.96	12	20대	311	25.92	12	시간	389	26.36
13	수학	91	35.64	13	문제	272	14.01	13	시작	355	11.72
14	여자	83	37.12	14	시간	272	15.91	14	모습	339	12.60
15	연락	82	21.50	15	상황	258	10.03	15	이유	313	15.56
16	시간	77	24.39	16	경찰	256	11.94	16	부모님	299	13.58
17	아빠	74	18.98	17	여성	256	13.24	17	연애	281	21.30
18	사진	70	35.81	18	대학	246	25.54	18	상황	274	9.75
19	얘기	70	19.60	19	연애	234	25.00	19	가치	268	12.18
20	학원	68	27.84	20	아빠	231	13.32	20	엄마	248	12.74
21	대학	60	20.29	21	마스크	229	35.61	21	회사	247	12.90
22	기분	56	24.77	22	이유	226	19.86	22	인생	241	17.62
23	이유	55	21.65	23	공부	199	13.42	23	경기	224	2.39
24	머리	54	22.75	24	학교	197	15.45	24	인간	224	11.59
25	고3	51	20.86	25	서울	187	13.17	25	주변	224	10.87
26	추천	51	30.48	26	내용	178	9.14	26	연락	222	11.47
27	남자애	50	19.67	27	얼굴	166	25.12	27	직장	213	12.71
28	스토리	47	19.43	28	전화	166	11.36	28	전화	208	12.36
29	노래	45	19.12	29	고민	162	13.98	29	남친	203	12.47
30	진심	44	18.53	30	관심	162	17.55	30	주변	224	10.81

III. Result

각 세대별 키워드 빈도 분석을 확인한 결과는 아래 Table 1과 같다. 먼저, 10대는 친구, 사람, 공부, 엄마, 학교 순으로 나타났으며, 20대는 사람, 친구, 생각, 남친, 엄마 순으로 나타났고, 30대는 사람, 생각, 친구, 남자, 결혼 순으로 나타났다. TF-IDF의 경우, 10대는 사람이 가장 높았으며, 친구, 남자, 공부 순으로 나타났다. 20대도 사람이 가장 높았으며, 친구, 남친남자, 여자 순으로 나타났고 30대는 남자가 가장 높았으며, 사람, 여자, 친구, 결혼 순으로 나타났다. 빈도 분석의 결과를 시각화하여 워드클라우드로 나타낸 결과는 Table 2와 같다.

Table 2. 세대별 게시판 키워드 워드클라우드



Table 3. 세대별 핵심 키워드의 연결중심성과 위세중심성

10대				20대				30대			
순위	keyword	연결중심성	위세중심성	순위	keyword	연결중심성	위세중심성	순위	keyword	연결중심성	위세중심성
1	친구	5166	1.000	1	생각	7102	1.000	1	사람	11210	1.000
2	사람	5030	0.958	2	얘기	4402	0.870	2	생각	9376	0.953
3	생각	3596	0.925	3	상황	4004	0.867	3	얘기	7892	0.934
4	학교	3172	0.880	4	마음	3896	0.856	4	문제	6576	0.856
5	공부	2880	0.708	5	문제	3834	0.839	5	시간	5614	0.832
6	엄마	2738	0.732	6	시간	3670	0.842	6	시작	5572	0.830
7	시간	2010	0.578	7	여자	3628	0.778	7	친구	5556	0.803
8	얼굴	1746	0.328	8	20대	3062	0.757	8	이유	5356	0.808
9	학원	1708	0.525	9	연락	3000	0.753	9	마음	5270	0.780
10	시험	1676	0.408	10	남자	2964	0.704	10	남자	4952	0.692
11	이유	1590	0.464	11	시작	2926	0.769	11	여자	4762	0.695
12	남자	1580	0.297	12	부모님	2898	0.740	12	결혼	4734	0.673
13	연락	1548	0.514	13	사람	2836	0.731	13	상황	4480	0.743
14	인스타	1494	0.298	14	이유	2828	0.740	14	주변	3784	0.693
15	사실	1486	0.538	15	엄마	2816	0.724	15	30대	3770	0.627
16	아빠	1478	0.451	16	가족	2712	0.731	16	모습	3758	0.698
17	수학	1468	0.306	17	상태	2704	0.709	17	이해	3608	0.688
18	얘기	1418	0.334	18	경찰	2690	0.599	18	과정	3486	0.656
19	남친	1398	0.316	19	사건	2438	0.654	19	처음	3250	0.666
20	가족	1378	0.707	20	친구	2256	0.660	20	과거	3174	0.621
21	부모님	1298	0.480	21	이해	2250	0.663	21	회사	3124	0.595
22	진심	1292	0.388	22	학교	2242	0.670	22	부모님	3046	0.559
23	머리	1246	0.333	23	처음	2210	0.667	23	상태	3032	0.636
24	기분	1244	0.345	24	내용	2206	0.640	24	자체	3030	0.638
25	주변	1242	0.353	25	공부	2180	0.615	25	인생	2986	0.585
26	짝남	1196	0.277	26	서울	2148	0.569	26	얼굴	2946	0.615
27	대학	1184	0.235	27	병원	2142	0.611	27	세상	2926	0.630
28	시작	1178	0.436	28	대학	2138	0.623	28	가치	2880	0.529
29	여자	1170	0.253	29	남친	2092	0.603	29	능력	2844	0.608
30	사진	1170	0.228	30	기분	2084	0.643	30	마지막	2770	0.606

마지막으로 연결중심성으로, 10대는 친구, 사람, 생각, 학교, 공부 순으로 나타났으며, 20대는 생각, 얘기, 상황, 마음, 문제 순으로, 30대는 사람, 생각, 얘기, 문제, 시간 순으로 나타났다. 위세중심성에서 10대는 친구, 사람, 생각, 학교, 엄마 순으로, 20대는 생각, 얘기, 상황, 마음, 시간 순으로, 30대는 사람, 생각, 얘기, 문제, 시간 순으로 연결 중심성과 유사한 모습을 보였다.

MZ 세대와 기성 세대와의 비교를 위하여, 네이트 판의 40, 50, 60 대의 게시판을 크롤링하여 비교 연구를 진행하고자 한다.

ACKNOWLEDGEMENT

본 논문(저서)은 교육부 및 한국연구재단 4단계 두뇌한국21 사업 (4단계 BK21 사업)으로부터 지원받은 연구임 이 논문 또는 저서는 2020년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2020S1A5B8103855)

REFERENCES

- [1] Univtomorrow. (2018). Trend MZ 2019 : Millennial-Z generation 5 marketing trends. Seoul : Hanbit Biz.
- [2] Lee, H. S., & Kim, J. H. (2021). Analysis of food consumption behavior due to COVID-19: Focusing on MZ generation. Journal of Digital Convergence, 19(3),

47-54.

- [3] Lee, J. Y., & Kim, S. I. (2022). A Study on Effect of Visual Elements of Cosmetic Package Design on Purchasing of MZ Generation, *Digital Convergence*, 20(2), 383-388.
- [4] Lee, S. G., & Lee, A. R. (2021). The Factors Influencing Value Awareness of Personalized Service and Intention to Use Smart Home : An Analysis of Differences between “Generation MZ” and “Generation X and Baby Boomers”. *Information Systems Review*. 23(3). 201-223