Improving UNIST Healthcare Center counseling Service using text mining

BAT57301 Service Systems Engineering and Management

Junmo Nam, Jingyu Jang

Abstract

As technology in data analyst field radically developed recently, it is not impossible to build model for unstructured data such as image, text. However, even though technology developed, some area such as healthcare rarely use techniques because of its characteristic such as extreme sensitive privacy, hard to define ontology and human feelings. Even with those obstacles, this research made test bed for automatic first-stage diagnose system for UNIST healthcare center which is expected to replace simple score-based survey. This tool analyzes UNIST students' worries based on answers that they did in survey session provided from UNIST health care center. First, descriptive analysis shows that there's similarity between survey data and part of NAVER data, tool is made based on two data set which is survey data from UNIST and NAVER Q&A data. There're three main function inside in initial tool, which is result from questions, result from answers and topic information table. Each information comes from combined model of word2vec and LDA with NAVER Q&A data and survey data. As tool developed by gaining further information from professionals in field for providing mode contents regard each topic, it is expected that students can freely get their own consultation without waiting by line.

Introduction

In [1]'s conclusion, recent news and opinions from faculties in healthcare center show that even though social awareness toward student's mental care increase, most of students are suffered by lack of treatment. This can be vary in what situation universities are in, but biggest obstacle measured in

healthcare center in UNIST is that lack of service provided compare to its demand(=students). This is mainly come from the fact that healthcare consulting requires professional skills and demand heavy duties to consultants for result that they provided to their customers.

On the other hand, there're many attempts to use chatbot in this area. For example, [2] investigated suitability of chatbots for a mental health intervention, specifically alcohol drinking habits assessment. It seems reasonable to provide chatbot service to patient in early stage so that service can follow up demands of students. Also, if auto-diagnose system for patients are well-designed in internet, students who have problem that hard to share with others fill more comfortable than communicate with someone, even if he/she is professional healthcare consultant. But unfortunately, even English based well-defined dictionary cannot make proper consultation toward patients. It is worse in case of Korea because the quality of dictionary is far less than English.

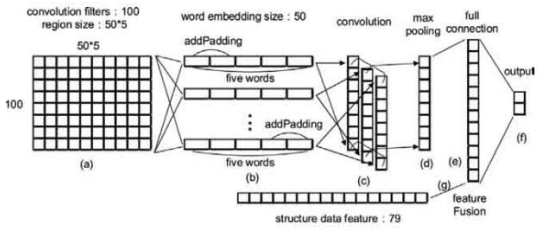
To overcome those obstacles, this research deals with quite narrow scope rather than common. As a result, this experiment will use survey data from UNIST healthcare center, which means that only UNIST student's case will be used. However, even though the range this research -UNIST students- will cover is smaller than usual healthcare service, the system should have its own scalability for the case when other problems occur in campus.

As chatbot rely on ontology of targeted area regardless of how deep its technology developed, one way to intensify in terms of scalability is that gather more text data related with that topic. Unfortunately, it is impossible to get more data related with such sensitive issue because of privacy, whether it is encrypted or not. In this research, Naver Q&A data which introduced in later part will take this role – enlarge fundamental ontology of chatbot.

Still it is doubtful about effectiveness of automatic consultation in healthcare area as such area demand sensitive treatment which machine cannot provide. So, to clarify the scope and function of service, next session will briefly review related articles.

Literature review

Word-to-vector is one of the widely used text mining technique that introduced first in [3]. The idea that word can be embedded into vector space and can compare differences between each other in more measurable manner changes the way that data analyst deal with text – which is infamous in its unstructured form. As method itself proved as it's more efficient and can interpret context of text compare to other simplified term based model such as TF-IDF combined with n-gram, word2vec model is widely used regardless of types of industry. One example that word2vec model implemented in healthcare industry is [4]. In [4], researchers tried to offer machine-learning flowline for predicting disease outbreak in disease frequent communities. The article used word2vec for gathering up more information that just normal(structured) data as usually medical data is incomplete to make highly accurate prediction. So they used CNN combined with word2vec to supplement those data.



CNN-based multimodal disease risk prediction (CNN-MDRP) algorithm

Another article is called 'Ask to doctor' service [5]. The purpose of system is quite similar to this paper's object. The purpose of article is that "automatically classify lay requests to an Internet medical expert forum using a combination of different text-mining strategies.". It also used word-based search for find analyzing start and synonym lists. One interesting thing that can be found in this article is that they use principal component analysis to reduce size of dimension and analyze how well each word predict the category of forum to recommend.

Table 3

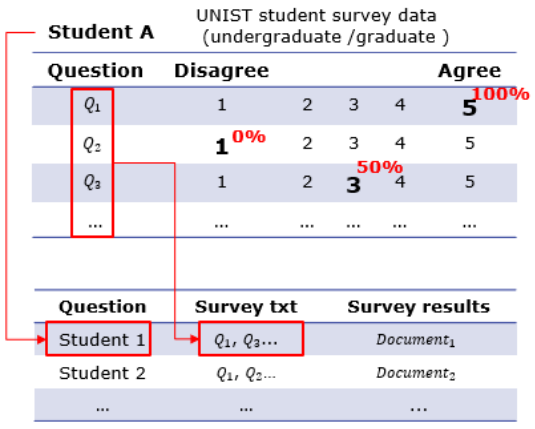
Most predictive words for the category "general information"

Word	Frequency, No. (%)		Cramer's V	*P*
	In "General Information"	In Other Categories		
X-chromosome	70 (13)	143 (31)	− 0.22	< .001
injection	17 (3)	68 (15)	− 0.21	< .001
utrogest	7 (1)	45 (10)	− 0.19	< .001
clomifen	32 (6)	82 (18)	− 0.19	< .001
prescribe	10 (2)	45 (10)	− 0.17	< .001
write	21 (4)	59 (13)	− 0.16	< .001
med	45 (8)	88 (19)	− 0.16	< .001
drug	24 (5)	59 (13)	− 0.15	< .001
pill	20 (4)	53 (12)	− 0.15	< .001
value	48 (9)	88 (19)	− 0.14	< .001
[places 11-50]				
fertile	36 (7)	44 (10)	0.12	< .001

It inspires us that if dictionary for text mining is well defined in pre-experiment phase, even so-called hardest area such as professional medical forum can be actually analyzed based on text mining techniques. And also, word2vec model is comparably insensitive toward number of dimension n, it is expected to provide more meaningful information on our case.

LDA(Latent Dirichlet Allocation), which is used to generate topic label for each question in this research, is one of the most famous algorithm in topic based text analysis. Related work can be found in recent text mining articles in journals, and [6] shows that LDA works on predicting students' behavior. In [6], they use students' text about each lesson without any given form such as survey questions or selective form, and use LDA to classify it and predict grade by SVM(Support Vector Machine). The result proves that LDA can efficiently divide text by group which is enough to contribute on model accuracy.

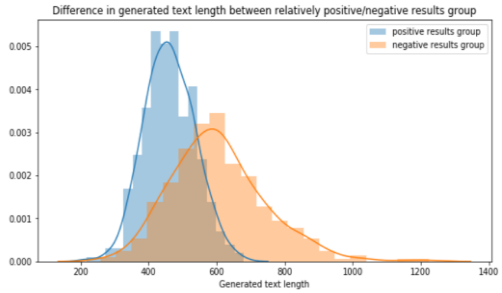
Analysis method



We conducted the analysis using survey data and Naver Q & A data. First, to extract text from survey

data, proceed as follows. We analyzed the results of the survey of freshmen students of UNIST in 2017 and 2018. In order to generate students' texts, four survey results were used such as (stress response inventory, PHQ-9, mini-international neuropsychiatric interview-plus) and GICC-15 (Global Interpersonal Communication Competence Scale). We gave each survey question 100% to 0% probability of the question text being generated as the student's text according to the degree of affirmation. For example, if student A answers 5 to the question "try not to feel my feelings to anyone", that is a strong affirmation, this question generates the student's text with a 100% probability. If student A answers 3 to this question (neutral answer), student A have 50 percent chance that the questions will be selected. The reason for this was to improve the variation of the students' texts. As a result, all students have their own documents based on their own answers and made the questionnaire.

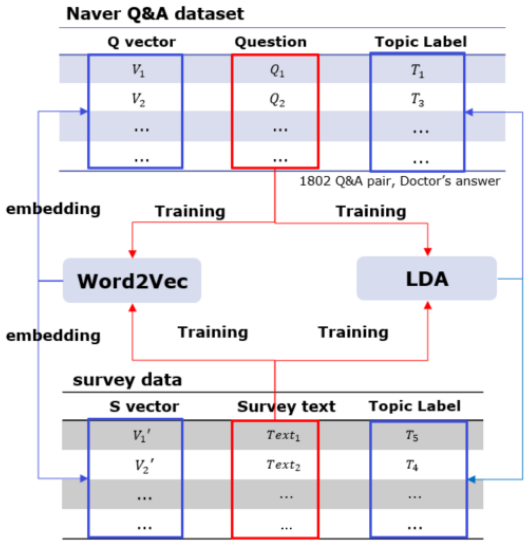
In the analysis, only 371 students were included in the analysis except for students who had no problem in all four of the survey. The graph below is a comparison of generated text lengths for groups that have some problems and without problems in all results. Because it is probabilistic, it can be seen that the generated text length of the group that received no problem in all surveys is shorter.



Second, Naver Q & A data was collected. In order to use only reliable answers related to mental health in Naver intellectuals, 2577 answers from some experts were collected - 'Kwon Suk-suk', 'Bae Seongbum', 'Shin Jae Hyun', 'Choi In-kwang', 'Lee Jae Won', 'Kim Yoon'. These areas of expertise are as follows. 1. gambling addiction and behavior addiction, corporate / worker mental health, interpersonal avoidance, anxiety / depressive disorder, dementia, sleep disorder, 2. deep psychological counseling, 3. psychological counseling, depression, (LGBTAIQ) counseling, family and couples treatment, psychiatric counseling, psychiatric counseling, sex therapy, sexual dysfunction, anxiety,

obsessive compulsive disorder, Psychotherapy, 5. Internet addiction, game addiction, smartphone addiction, ADHD, adult ADHD, cognitive disorder, brain stimulation, 6. Depression, anxiety, vase, cognitive decline, interpersonal problems, dementia, ADHD, Disability / restless leg syndrome, obsessive compulsive disorder, amyotrophic lateral sclerosis, bipolar disorder, and fatigue.

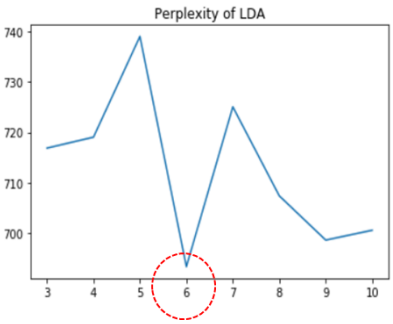
Special characters and unnecessary spaces were removed for preprocessing of collected text, and Mecab module of eunjeon package was used for noun extraction. After extracting the nouns of a letter difficult to grasp the meaning, it was found that there was a little difference between the question of the survey. We removed NAVER Questions with nouns such as 'Dementia', 'Youth', 'SAT', 'Menopause', 'Superior', 'divorce', 'disability', 'ghost', 'baby', 'husband', 'groom'. The title of the question extracted only by nouns was added to the front of the question, and the question of less than 10 characters was thought to be inadequate length, so those questions were removed. Finally, a total of 1802 questions and answers were used for analysis.



For the LDA, a term frequency matrix was constructed by combining NAVER Question, Title, and Survey texts. TF matrices were created with only question and survey texts in this term frequency matrix. The total number of unique words was 6623. We then used the Survey Text data and NAVER Question data to model the LDA model and the Word2vec model for modeling. The topic that can be obtained from the LDA learning result is designated as the label of the document. In the case of the Word2vec

model, the text is vectorized by learning NAVER Question data and Survey text data.

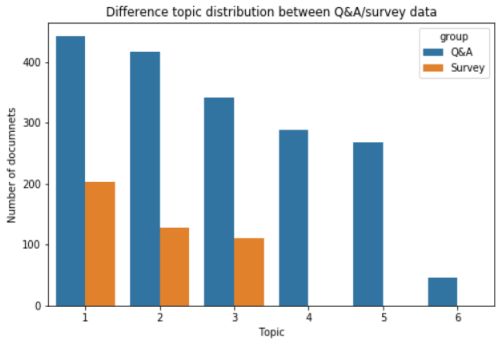
The reason for this structure is that the survey text and NAVER Q & A data have different context. Therefore, we use predictive modeling to confirm the feasibility of applying NAVER Q & A data for UNIST students. We considered that if the prediction performance was high enough, two different data sets could be used together.



The LDA was designed by specifying 6 topics with minimum perplexity score. The smaller the value is, the better the learning is because it means that the topic model reflects the actual document results well.

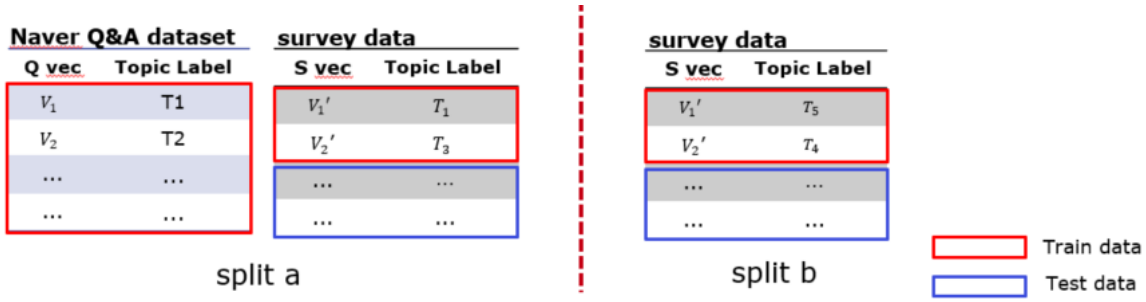
```
Topic 1:
사람 친구 생각 엄마 학교 부모 가족 남자 성격 아빠
Topic 2:
생각 정신 정도 병원 우울증 치료 증상 불안 사람 때문
Topic 3:
게임 중독 입원 시간 컴퓨터 알콜 인터넷 본인 하루 치료
Topic 4:
대화 상대 자신 사람 집중 생각 친구 관심 고갯짓 반말
Topic 5:
결벽증 부작용 거식증 페니 사워 보험 정신 적용 생리 터치
Topic 6:
강박 공포증 확인 복용 벌레 사고 다리 시선 현상 노래
```

The results of the six topics extracted from the LDA were as follows. In Topic 1, words such as people, friends, thoughts, mothers, and schools were predominantly predictive of relationship problems. In Topic 2, words related to depression or unrest, such as thoughts, mental, degree, hospital, depression, came out. In Topic 3, words related to addiction such as game, addiction, hospitalization, time, computer, alcoholism or computer addiction came out. In Topic 4, words related to the communication problem such as conversation, opponent, self, person, concentration etc. mainly came out. In Topic 5, words related to mysophobia and anorexia, such as dizziness, side effects, and anorexia, came out. In Topic 6, words related to obsession and fear such as obsessive-compulsive fright.



The distribution of topic in each data set with generated topic can be seen that Q & A data is relatively evenly distributed in the topic label except 6 topic label. In Survey text data, it shows that there is data only in 1,2,3 topics. This is expected result because the variation of text is smaller than Question. Question 1, 2, 3, 4, 5 and 6 had 442, 417, 342, 289, 267 and 45 labels, respectively. The 1,2,3 labels of the survey text were 202, 127, and 111, respectively.

We used the skip gram model to training the Word2vec model. We set the vector size to 300 and the window size to 10. We used the NAVER Question and Survey texts to learn the Word2vec model, and we obtained vectors of words in the form of projecting each word into the model. Each document vector was obtained by averaging the word vectors. A total of 2242 document vectors of 300 dimensions became final data sets.



Finally, train and test sets were divided to predict the topic label. In the first split a, 50% of the NAVER Question and 50% of the survey text were used as the train set and 50% of the remaining survey text was used as the test set. In the second split b, only survey texts were used, 70% train sets, and 30% test sets. The reason for separating the datasets in two ways is to find out the impact of the NAVER Q & A data in survey results prediction.

Random forest and Light GBM models were used as prediction model algorithms. Both models are ensemble models and are known to exhibit stable and excellent performance.

Results

Random Forest

test fit report

	precision	recall	f1-score	support
1	0.59	0.85	0.70	101
2	0.61	0.30	0.40	64
3	0.86	0.69	0.77	55
avg / total	0.67	0.65	0.63	220

test fit report

	precision	recall	f1-score	support
1	0.67	0.95	0.79	61
2	0.67	0.32	0.43	38
3	0.86	0.73	0.79	33
avg / total	0.72	0.71	0.68	132

Light GBM

test fit report

	precision	recall	f1-score	support
1	0.66	0.87	0.75	101
2	0.58	0.39	0.47	64
3	0.86	0.69	0.77	55
avg / total	0.69	0.69	0.67	220

test fit report

	precision	recall	f1-score	support
1	0.74	0.92	0.82	61
2	0.64	0.42	0.51	38
3	0.77	0.73	0.75	33
avg / total	0.72	0.73	0.71	132

Spli-a Split-b

In Split a, Randomforest and Light GBM were 67 and 69% respectively. In Split b, the accuracy was about 72% in both algorithms. It is important to make sure that the accuracy of the model is quite stable even if the Q & A data are trained together. Based on these results, we conclude that when a new input comes in the form of a survey or text, it is valid to suggest an appropriate answer using NAVER Answer.

Application

We made the service using the preprocessing module and the LDA model used in the front, and we selected R shiny as the final service type. R shiny has the advantage of interactive plot visualization by

receiving user input. In the above picture, when the user's question related to the study is input, the preprocessing is completed in the existing preprocessing module and the most similar topic is found in the learned LDA model. And using the text in the topic to show the wordcloud. The goal was to alleviate the psychological burden of the users by showing that there are many people who are experiencing similar problems with the problems they are experiencing.

In the same way, the above figure provides the answers of the same Topic doctors in wordcloud. This allowed users to obtain simple but helpful information.

Finally, this graph shows the main words of the topic of the user and shows a lot of other similar to the situation that you are experiencing.

Discussion

This project aims to improve the self-diagnosis service of UNIST Healthcare Center using NAVER

Q&A and UNIST freshman survey. This service introduces similar cases of mental health that UNIST students may experience, such as stress, depression, suicide, etc., and allows them to present their answers in wordcloud on related topics.

We have created some services using NAVER Q&A data because of the difficulty of consulting the students collected in text form and the lack of answers from the doctors. However, if actual data of UNIST students accumulated in the future, We can expect to provide better service.

Reference

[1] Eisenberg, Daniel, Ezra Golberstein, and Sarah E Gollust. “Help-Seeking and Access to Mental Health Care in a University Student Population.” Medical Care 45, no. 7 (2007): 594–601.

[2] Elmasri, Danielle, and Anthony Maeder. “A Conversational Agent for an Online Mental Health Intervention.” In International Conference on Brain and Health Informatics, 243–251. Springer, 2016.

[3] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. “Efficient Estimation of Word Representations in Vector Space.” ArXiv Preprint ArXiv:1301.3781, 2013.

[4] Chen, Min, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. “Disease Prediction by Machine Learning over Big Data from Healthcare Communities.” IEEE Access 5 (2017): 8869–8879.

[5] Himmel, Wolfgang, Ulrich Reincke, and Hans Wilhelm Michelmann. “Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums.” Journal of Medical Internet Research 11, no. 3 (2009).

[6] Sorour, Shaymaa E, Kazumasa Goda, and Tsunenori Mine. “Estimation of Student Performance by Considering Consecutive Lessons.” In Advanced Applied Informatics (IIAI-AAI), 2015 IIAI 4th International Congress On, 121–126. IEEE, 2015.