

환자의 주관적 증상 텍스트에 대한 진료과목 분류 모델 구축

Classification Modeling for Predicting Medical Subjects using Patients' Subjective Symptom Text

이서희 · 강주영*

아주대학교 경영대학 e-business학과

요 약

의료 인공지능 분야에서 의사의 판단에 도움을 줄 수 있는 질환 예측 및 분류 알고리즘에 대해선 많은 연구가 이뤄져왔지만, 의료 소비자의 정보 획득과 판단에 도움을 줄 수 있는 인공지능에 대해선 상대적으로 관심이 적다. 네이버 지식인에 지난 1년 간 자신의 증상엔 어떤 병원을 가야할 지 질문하는 질문 건수만 해도 15만 건이 넘는다는 사실은 의료소비자들에게 적합한 의료정보의 제공이 필요하다는 반증이기도 하다. 따라서 본 연구에선 의료소비자들이 자신의 증상에 대한 진료과목을 선택하는데 도움을 줄 수 있도록 네이버 지식인에서 환자들이 직접 서술한 증상 텍스트를 수집하여 8개 진료과목을 분류하는 분류모델을 구축했다.

우선 환자의 주관어 개입된 데이터의 타당성과 객관성을 확보하기 위해 객관적 증상 텍스트(서울응급의료 정보센터에서 정리한 진료과목 별 주요 질환 증상)와 주관적 증상 텍스트(지식인 데이터) 간 유사도 측정을 수행하였다. 유사도 측정 결과, 두 텍스트가 동일한 진료과목의 증상일 경우 상이한 진료과목의 증상 텍스트에 비해 상대적으로 높은 유사성을 가진다는 것을 입증했다. 상기 절차를 따라 타당성을 확보한 주관적 증상 텍스트를 대상으로 릿지회귀모델을 사용하여 분류모델을 구축한 결과 0.73의 정확도를 확보할 수 있었다.

■ 중심어 : 진료과목 분류, 머신러닝, 유사도 분석, 주관적 증상

Abstract

In the field of medical artificial intelligence, there have been a lot of researches on disease prediction and classification algorithms that can help doctors judge, but relatively less interested in artificial intelligence that can help medical consumers acquire and judge information. The fact that more than 150,000 questions have been asked about which hospital to go over the past year in NAVER portal will be a testament to the need to provide medical information suitable for medical consumers. Therefore, in this study, we wanted to establish a classification model that classifies 8 medical subjects for symptom text directly described by patients which was collected from NAVER portal to help consumers choose appropriate medical subjects for their symptoms.

In order to ensure the validity of the data involving patients' subject matter, we conducted similarity measurements between objective symptom text (typical symptoms by medical subjects organized by the Seoul Emergency Medical Information Center) and subjective symptoms (NAVER data). Similarity measurements demonstrated that if the two texts were symptoms of the same medical subject, they had relatively higher similarity than symptomatic texts from

different medical subjects. Following the above procedure, the classification model was constructed using a ridge regression model for subjective symptom text that obtained validity, resulting in an accuracy of 0.73.

■ Keyword : Medical subjects classification, machine learning, similarity analysis, subjective symptoms

I. 서론

개인의 의료기록이나 데이터를 대상으로 질병을 예측하고 진단하는 질병 진단 알고리즘에 대해서 현재까지 많은 연구들이 있어왔고 지금도 해당 분야에 대해 활발한 연구가 진행 중이다. 환자의 MRI 사진을 분석하여 알츠하이머 병을 분류하는 연구[10]에서부터 CT 이미지를 분석하여 COVID-19를 진단하는 AI[8]까지 의료 AI기술은 다양한 질환을 진단하고 예측하기 위해 발전하고 있다. 해당 기술들은 의사가 자료들을 분석하고 판독하는데 소요되는 시간을 줄이고 보다 정확한 판단을 내릴 수 있도록 보조하는 용도로 사용되고 있다. 이렇듯 의료 AI는 의사의 판단에 도움을 주는 방향으로 발전해가고 있지만 의료소비자들의 판단과 정보획득에 도움을 주는 AI에 대해서는 상대적으로 관심과 관련 연구가 적다.

본 연구는 의료 소비자들이 병원을 가고 진단받기까지의 과정에서 어떠한 증상이 나타났을 때 어떤 진료과목을 다루는 병원에 갈지 찾아보고 선택하는 과정에 도움을 줄 수 있는 AI를 연구하는데 목적이 있다. 네이버 지식인에 “어느 병원”을 키워드로 검색해보았을 때 2020년 1월부터 2021년 1월까지 1년 사이의 기간 동안 검색되는 결과 건수만 해도 15만 건이 넘는다. 이는 그만큼 의료소비자들이 병원을 선택하는 데 혼란의 여지가 있고 증상에 대한 적절한 진료과목에 대한 판단에 어려움이 있다는 반증일 것이다. 따라서 본 연구를 통해 의료소비자들이 자

신의 증상에 대해 서술하면 그에 적합한 진료과목이 무엇인지 분류하는 모델을 구축하고자 하였다.

본 연구에선 환자가 직접 서술한 증상을 대상으로 분류모델을 구축하기에 앞서 환자의 주관성이 많이 함유된 데이터에 대한 검증이 필요했다. 나이가 어릴수록 진성안증후군으로 인한 주관적 증증도가 높다[2]거나 주관적 건강상태가 나쁜 여성일수록 강박 증상을 더 많이 호소하는 경향[1]이 있다는 연구들을 미루어보았을 때 환자가 호소하는 증상은 나이, 주관적 심리상태 등에 따라 일반적으로 서술되는 질병의 증상과 다소 차이가 있을 수 있기 때문이다. 따라서 본 연구에서 분류 모델을 구축하기 전, 환자가 주관적으로 호소하는 증상텍스트와 진료과목 별로 일반화되어 나름의 객관성을 갖춘 증상 텍스트 간의 유사도를 분석하였다. 해당 분석을 통해 두 텍스트에 차이가 없는지 살펴봄으로써 주관적 증상 데이터의 타당성과 객관성을 확보하고자 하였다. 본 논문에선 용어의 편의를 위해 분류 모델의 데이터로 사용된 주관성이 함유된 증상 데이터를 ‘주관적 증상’으로, 각 진료과목 별 주요 질환 의 일반적인 증상으로 정리된 텍스트 데이터를 ‘객관적 증상’으로 이름 붙여 사용하였다.

상기 유사도 분석을 거쳐 타당성을 확보한 주관적 증상 데이터를 대상으로 진료과목 별 증상 내용에 차이가 있는지를 tf-idf 기반의 키워드 분석으로 확인해보았다. 마지막으로 해당 데이터를 대상으로 분류 모델을 구축함으로써 의료소비자들의 판단에 도움이 될 수 있는 인공지능을

구축하는데 기여하고자 하였다.

II. 관련 선행연구

의료분야에서 자연어처리를 이용한 연구로는 환자의 임상 전자 기록에서 하지 말초 동맥 질환과 관련된 개념을 추출하여 해당 질환을 자동으로 진단하는 연구[6]와 전자 의료 기록에서 환자가 뇌 동맥류 질환을 가질 확률을 추정하는 연구[9]와 같이 전자 의료기록을 대상으로 특정 질환을 예측하는 연구들이 선행되어왔다. 이 뿐만 아니라 구글에서는 자연어처리 딥러닝 기술을 활용하여 전자의료기록과 의료진의 진료노트를 대상으로 입원 환자의 사망 위험도와 퇴원 후 30일 내 재입원 확률 등을 예측하는 모델[13]을 연구하기도 했다.

위와 같이 의료 분야에서는 자연어처리를 통해 병원에서 나온 전자 진료기록 등과 같은 데이터를 학습하여 의사의 판단과 업무에 도움을 주는 인공지능 연구가 많이 이뤄져왔다. 그러나 본 연구에서는 자연어처리를 통해 환자에게서 나온 데이터만을 사용하여 환자의 판단에 도움을 줄 수 있는 진료과목 분류 모델을 구축하는 연구를 수행하였다.

진료과목 예측과 관련한 연구로는 국민건강정보데이터를 활용하여 진료과목을 분류하는 모델을 구축한 연구[12]가 있다. 해당 연구에선 국민건강정보데이터의 건강검진 정보와 진료내역 정보를 데이터로 내과, 소아청소년과, 정형외과, 산부인과, 이비인후과 총 5개의 진료과목에 대한 진료과목 예측 모델링을 구축하였고 SVM(Support Vector Machine)모델로 평균 80%이상의 정확도를 확보하였다.

해당 선행 연구는 요양일수, 총 처방일수, 콜레스테롤 수치 등 대부분 수치데이터를 대상으로 분류 모델을 구축하였으나, 본 연구에선 수치데이터가 아닌 환자가 직접 서술한 증상 텍스

트 데이터를 대상으로 8개 진료과목에 대한 진료과목 분류 모델을 구축하였다는 점에서 선행 연구와 차이가 있다.

III. 연구 방법

연구는 <Figure 1>과 같은 프로세스로 진행하였다.

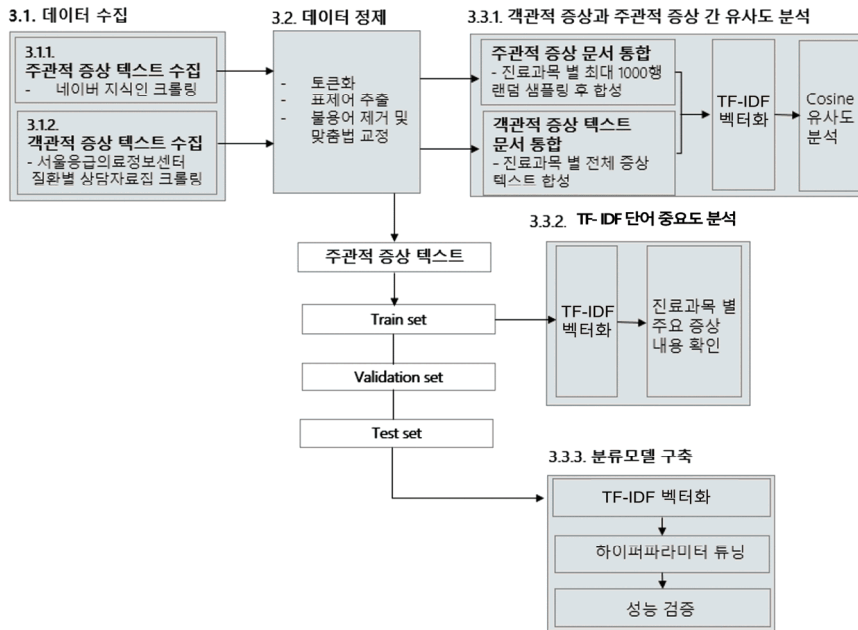
3.1 자료 수집

본 분석에서 활용한 데이터는 총 두 가지이다. 첫 번째는 분류모델링을 학습하기 위해 사용된 텍스트 데이터로 네이버 지식인 페이지를 크롤링한 데이터이다. 두 번째 데이터는 첫 번째 데이터의 타당성을 확보하기 위해 비교대상이 되는 데이터로, 서울응급의료정보센터에서 작성한 진료과목 별 증상과 전문의 의견 등을 담은 질환별 상담자료집을 크롤링하여 사용하였다.

3.1.1 주관적 증상 텍스트 수집

첫 번째 데이터를 구축하기 위해 파이썬의 Selenium과 BeautifulSoup패키지를 사용하였다. Selenium패키지는 다양한 웹브라우저를 자동 조작할 수 있는 도구로, 해당 패키지를 사용해 Chrome 브라우저에서 자동으로 원하는 내용을 검색할 수 있도록 조작하였다. BeautifulSoup는 html을 파싱하여 데이터를 추출할 수 있도록 돕는 패키지로 해당 패키지를 사용해 html에서 원하는 부분의 데이터들을 수집할 수 있었다.

2010년 1월부터 2021년 5월 7일까지의 기간 동안 네이버 지식인에서 “어느 병원”, “어떤 병원”, “어디 병원”, “무슨 병원”, “병원 어디”, “어느 진료과”, “어떤 진료과”, “어디 진료과”, “무슨 진료과”, “진료과 어디”로 총 10개의 키워드와 정확히 일치하는 키워드를 포함하여 업로드



<Figure 1> 연구 프로세스

된 질문 중에서 의사 답변이 달린 데이터들만 검색결과로 도출될 수 있도록 필터링 하였다. 크롤링된 데이터를 대상으로 <Figure 2>와 같이 질문 텍스트는 증상데이터로, 의사 답변 텍스트는 답변 중 ‘신경과’와 같은 진료과목 부분만 추출하여 타겟 데이터로 사용하였다. 추출한 진료과목은 내과, 비뇨기과, 산부인과, 신경과, 안과,

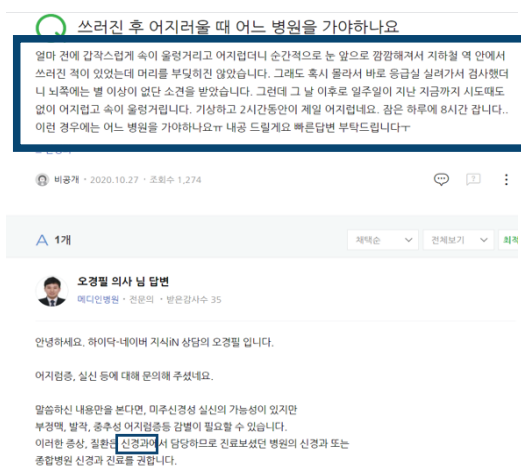
외과, 이비인후과, 정형외과 총 8개 진료과목으로, 데이터 중 타겟 데이터가 8개 진료과목에 속하지 않거나 증상 내용이 중복되는 데이터들은 모두 삭제하였다. 결과적으로 9543건의 데이터를 확보할 수 있었다.

3.1.2 객관적 증상 텍스트 수집

두 번째 데이터인 객관적 증상 데이터는 서울 응급의료정보센터에서 제작한 질환 별 상담자료집[3]을 pdf 크롤링 패키지인 tika를 사용하여 수집하였다. 첫 번째 데이터와 동일한 8개 진료과목을 대상으로, 진료과목 별 자주 진단되는 증상이 정리된 페이지의 문항들을 크롤링하였다. 결과적으로 총 4263건의 데이터를 확보할 수 있었다.

3.2 데이터 정제

확보한 데이터에서 특수문자나 반복되는 자음 등을 제거한 후 한글 맞춤법 검사기인 hans-



<Figure 2> 주관적 증상 데이터 예시

pell라이브리리를 통해 맞춤법을 교정하였다. 문장 품사 중 명사, 형용사, 동사만 추출하여 분석에 사용하였고 동사와 형용사의 원형을 대상으로 분석하기 위해 표제어 추출을 수행하였다. 그 후 ‘진료과목’, ‘어떻다’와 같이 분석에 불필요한 단어들은 불용어로 제거하였으며 단어의 길이가 2개 이상인 단어들만 추출하여 분석에 사용하였다.

3.3 자료 분석

우선 수집한 주관적 증상 데이터와 객관적 증상 데이터를 대상으로 진료과목 별 유사도 분석을 하였다. 다음으로 분류 모델에 사용할 주관적 증상 텍스트 데이터들을 tf-idf 벡터화한 후 진료과목 별 단어들을 중요도 순으로 나열하였다. 나열된 단어들을 살펴봄으로써 진료과목 별 중요 단어와 내용 차이를 파악하였다.

3.3.1 객관적 증상과 주관적 증상 간 유사도 분석

객관적 증상과 주관적 증상 간 유사도를 측정하기 위해 문서 간 코사인 유사도(cosine similarity)를 도출하여 분석하였다. 코사인 유사도는 두 벡터의 각도가 일치하여 거의 동일하면 1 이, 각도가 커서 유사하지 않으면 0이 도출된다.

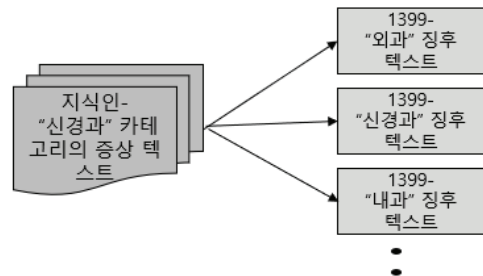
$$\frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

〈Figure 3〉 코사인 유사도 산출 공식

본 연구에선 객관적 증상과 주관적 증상 데이터를 함께 tf-idf 벡터화한 뒤 벡터 간 코사인 유사도를 측정[14]함으로써 문서 유사도를 분석하고자 했다.

우선 두 데이터의 진료과목 별 유사도를 분석하기 위해 각 데이터에서 진료과목 별 증상을

통합하여 하나의 문서를 생성하였다. 객관적 증상의 경우 데이터 수가 많지 않아 진료과목 별 모든 데이터를 합쳐 하나의 문서로 생성하였고, 주관적 증상의 경우 최대 1000개의 행을 랜덤으로 추출하여 하나의 문서로 통합하였다. 그 후 문서들을 벡터화한 후 주관적 증상과 객관적 증상 문서들의 유사도를 <Figure4>와 같은 방법으로 측정하였다.



〈Figure 4〉 문서 간 유사도 측정 방법

코사인 유사도는 단어의 의미적 유사도는 알 수 없기 때문에 같은 의미라도 용어가 다르면 수치가 떨어질 수밖에 없다는 단점이 있다. 본 연구에선 객관적 증상과 주관적 증상 텍스트에 대한 용어 표준화 과정을 거치지 않았기에 코사인 유사도의 절대적인 수치로 유사도를 측정하기엔 적합하지 않을 수 있다.

따라서 본 분석에선 특정 진료과목에 대한 주관적 증상과 모든 진료과목에 대한 객관적 증상들의 유사도를 측정했을 때 서로 동일한 진료과목의 증상 텍스트 유사도가 상이한 진료과목의 유사도보다 상대적으로 더 큰 값을 도출하는지 확인해보고자 했다. 만일 동일한 진료과목의 증상 텍스트 유사도가 다른 진료과목과의 유사도보다 더 높은 값을 가진다면, 적어도 본 연구에서 사용되는 데이터에서는 주관적 증상과 객관적 증상 텍스트가 서로 유사성을 지닌다고 추정할 수 있다.

3.3.2 TF-IDF 단어 중요도 분석

주관적 증상 데이터를 대상으로 모델을 학습하기 전에 데이터에 대한 간단한 탐색적 분석을 수행하고자 했다. 모델에 데이터를 학습시키기 전, 정제가 끝난 텍스트 데이터를 수치화하는 과정이 필요한데 본 연구에선 문서에서 단어의 중요도를 잘 표현하는 TF-IDF(Term Frequency - Inverse Document Frequency) 벡터화를 통해 데이터를 수치화하였다. tf-idf는 문서에서 나타난 단어의 빈도를 그 단어가 나타난 문서의 수로 나뉜 값으로 해당 단어가 문서에서 어느 정도의 중요도를 지니는 지를 반영할 수 있는 벡터 표현이다.

우선 주관적 증상 데이터 셋을 0.64 : 0.16 : 0.2 비율의 학습 데이터와 검정데이터, 테스트 데이터로 나누었다. 먼저 학습 데이터를 tf-idf 벡터화하고 해당 벡터를 진료과목 별로 나눠 중요도 순으로 상위 20개 단어를 나열하였다. 이를 통해 각 진료과목 마다 어떤 단어가 중요하고 중요 단어의 양상에 진료과목 별로 차이가 충분히 존재하는지 확인해보았다.

3.3.3 분류 모델 구축

벡터화한 데이터를 대상으로 파이썬의 Scikit-Learn 패키지를 활용해 나이브베이지 분류(Multinomial Naïve Bayes), 랜덤포레스트 분류(RandomForest Classifier), 서포트벡터머신(Linear Support Vector Classifier), 로지스틱 회귀(Logistic Regression) 모델을 매개변수 값을 모두 초기 값으로 두고 학습시켰다. 그 후 네 모델을 정확도 기준으로 평가한 결과, 로지스틱 회귀 모델의 성능이 0.69로 가장 높았다. 과적합이 존재할 수 있는 로지스틱 회귀모형에서 각 특성에 대한 계수들이 지나치게 높아지는 것을 방지하기 위해 회귀모델을 정규화한 릿지 회귀모델(Ridge Classifier)을 최종 모델로 선택하였다. 릿지 회귀 모델에서 정규화의 정도를 조절하는 al-

pha 매개변수를 조정함으로써 검정 데이터에 대한 성능이 최대가 되도록 설정하였다.

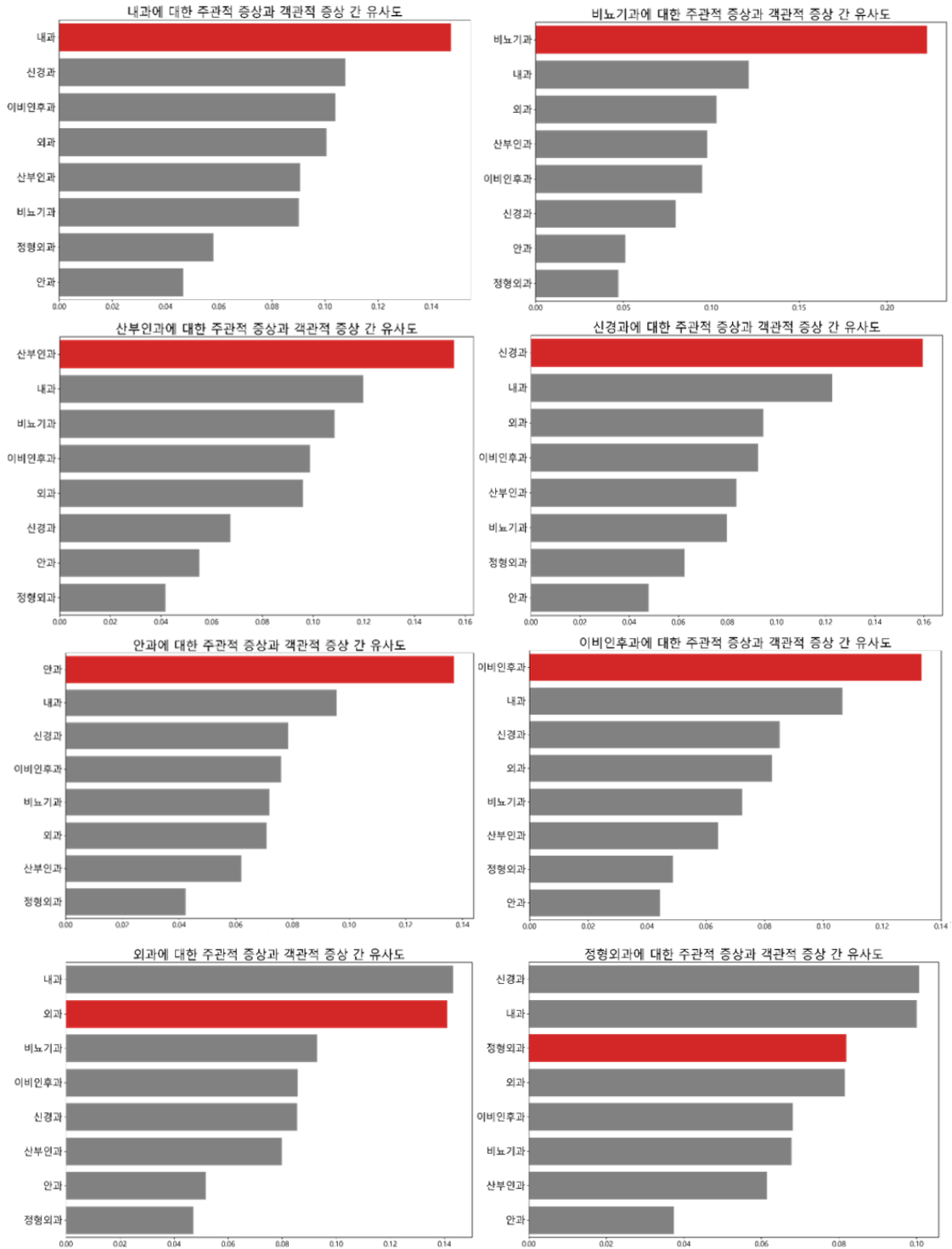
IV. 연구 결과

4.1 주관적 증상과 객관적 증상 간 유사도 분석 결과

문서 유사도 분석 시 비교하는 대상은 특정 진료과목에 대한 주관적 증상 문서이며, 비교되는 대상들은 8개 진료과목에 대한 객관적 증상 문서들로 설정하여 각각의 코사인 유사도 값을 구해 막대그래프로 시각화하였다. <Figure 5>은 그래프 제목에 명시되어있는 진료과목의 주관적 증상에 대한 8개 객관적 증상 문서의 유사도를 나타내는 그래프이다.

<Figure 5>를 보면 용어의 표준화가 이뤄지지 않아 전반적으로 절대적인 유사도 수치는 낮지만, 8개 객관적 증상 문서들 모두 주관적 증상과 동일한 진료과목의 유사도 순위가 나며 7개 진료과목의 유사도와 비교했을 때 상위 3위 내에 들어오는 것을 볼 수 있다. 즉 환자들이 호소하는 증상 텍스트에 주관성이 다소 포함되어 있다 할지라도 증상을 호소하기 위해 사용하는 단어나 구문은 일반적으로 진료되는 질환의 증상과 유사하다고 할 수 있다. 주관적 증상과의 유사도가 가장 낮은 3위를 기록한 정형외과의 경우 객관적 증상 문서에서 정형외과에서 자주 진료되는 질환을 설명할 때 ‘반월’, ‘경골’ 등 일반 사람들이 자신의 증상을 설명할 때 잘 사용하지 않는 전문적 용어가 더 많았을 뿐 만 아니라 ‘신경 손상’, ‘통증’과 같은 단어들이 신경과의 단어들과 상당한 유사성을 보이기 때문에 이러한 결과를 도출한 것으로 보인다.

본 분석을 통해 전반적으로 지식인에서 추출한 주관적 증상이 일반적으로 진료되는 질환의 주요 증상과 상대적으로 유사함을 보였다. 따라



<Figure 5> 주관적 증상과 객관적 증상 간 유사도 분석 결과

서 해당 데이터가 진료과목 분류 모델링을 구축 하는데 적합하다고 판단하였다.

4.2 TF-IDF 단어 중요도 분석 결과

주관적 증상 데이터에서 진료과목 별로 주요 키워드 분포에서 차이가 있는지 탐색해보고자 진료과목 별 단어 중요도를 살펴보았다. 분류 모델에 입력할 텍스트 데이터는 tf-idf 수치 벡터 형태로 변환되어 들어간다. tf-idf는 문서 내에서 해당 단어가 얼마나 중요한지를 나타내기 때문에 수치가 높을수록 문서에서 더욱 중요한 단어 라고 할 수 있다. 진료과목 별로 어떤 단어가 중요한지 보고자 진료과목 별 상위 20개의 tf-idf 벡터들을 출력하였고 해당 결과를 <Table 1>에서 볼 수 있다.

진료과목 별 주요 단어들을 살펴보면 외과는 피부나 특정 신체부위에 무엇인가가 만져질 때, 비뇨기과는 소변, 또는 비뇨기에 문제가 있을 때, 신경과는 머리카락이나 특정 신체부위에 통증이

있을 때, 내과는 가슴이나 머리카락이 답답하거나 이상이 있을 때, 이비인후과는 귀와 코에 문제가 있을 때, 안과는 안과 수술 또는 눈에 이상이 있을 때, 정형외과는 다리, 무릎, 발톱과 같은 신체 부위에 이상이 있을 때, 산부인과는 생리기 관에 문제가 있을 때 주로 방문한다는 것을 파악할 수 있다. ‘머리카락’, ‘어지럽다’ 등 중요한 단어가 겹치는 과목들도 몇몇 볼 수 있는데, 이는 질환마다 복합적인 증상이 함께 연관되어 나타날 수 있기 때문[4][15]인 것으로 보인다. 대체적으로 몇 가지 겹치는 단어들을 제외하고는 진료과목 별 주요 단어들에 차이를 확인할 수 있다.

4.3 분류 모델 구축 결과

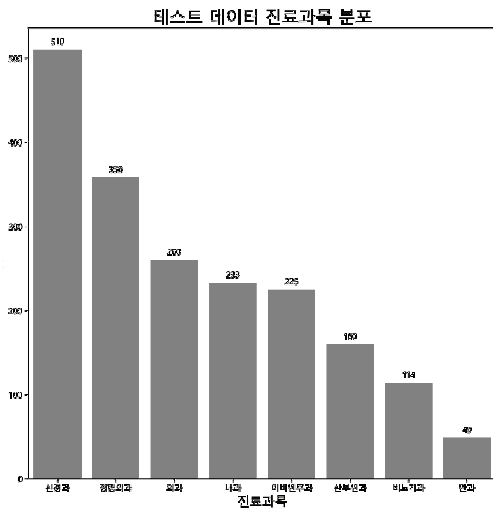
총 9543건의 데이터에서 학습데이터는 0.68의 비율로 6108건, 검증 데이터는 0.16의 비율로 1526건, 테스트 데이터는 1909건이며 전체 진료과목 타겟의 비율을 반영할 수 있도록 진료과목 기준으로 층화 추출하였다. 8개 진료과목을 분

<Table 1> 진료과목별 TF-IDF 단어 중요도

신경과	정형외과	외과	내과	이비인후과	산부인과	비뇨기과	안과
머리카락	다리	생기다	가슴	생기다	소음	소변	라섹
두통	허리	만지다	검사	어지럽다	수술	비뇨기	라식
왼쪽	왼쪽	멍울	자다	부다	생리	검사	수술
오른쪽	어깨	겨드랑이	기침	소리	늘어나다	방광염	시력
자다	무릎	수술	심하다	자다	늘어지다	화장실	다래끼
어지럽다	오른쪽	가슴	쉬다	수술	검사	오줌	각막
심하다	손가락	오른쪽	왼쪽	머리카락	소변	고환	쌍꺼풀
떨리다	발톱	왼쪽	머리카락	비염	자궁	자주	검사
허리	없다	누르다	갑자기	왼쪽	질염	수술	좋다
신경	발가락	피부	답답하다	목소리	생리통	자다	오른쪽
일어나다	운동	잡히다	어지럽다	어지럼증	좋다	성기	왼쪽
다리	심하다	항문	오른쪽	오른쪽	임신	여자	보이다
저리다	손목	유방	일어나다	검사	불편하다	혈뇨	부다
갑자기	발목	유두	자주	귀가	관계	싸다	떨리다
검사	수술	검사	좋다	심하다	여자	방광	렌즈
괜찮다	근육	크기	심장	갑자기	비용	질문	자다
없다	자다	커지다	두통	삼키다	골반	남자	한쪽
어깨	골반	튀어나오다	생기다	입안	성관계	잔노감	심하다
좋다	생기다	딱딱하다	괜찮다	안쪽	심하다	요도	눈꺼풀
저림	허벅지	부다	가끔	만지다	자다	사타구니	머리카락

류하는 모델을 Sckit-learn의 Ridgeclassifier를 사용하여 구축하였고 정규화의 강도를 조절하는 매개변수인 alpha값은 3.8로 검증 데이터에서 최고 성능이 도출된 값으로 설정하였다.

다음 <Figure 6>는 진료과목 별 테스트 데이터에서 진료과목의 분포를 보여주는 그래프이다.



<Figure 6> 테스트 데이터 진료과목 분포

“신경과”로 분류되어있는 데이터가 510건으로 가장 많고 “안과”는 49건에 불과했다. 이렇듯 데이터에 클래스 불균형이 존재할 경우 모델을 정확하게 평가하기에 정확도(accuracy)는 크게 의미가 없을 수 있다[5]. 따라서 해당 모델을 평가하기 위해 정확도 이외에도 정밀도(precision), 재현율(recall), F1-score를 추가적으로 활용하고자 했다. 만일 외과 분류모델에 대한 정밀도가 0.5라면 모델이 외과라고 예측한 데이터들 중 실제 진료과목이 외과인 데이터의 비율이 0.5라는 의미이다. 재현율이 0.5라면 실제로 외과인 데이터 260개 중에 모델이 130개는 외과라고 제대로 예측했다는 의미이다. F1-score는 이 둘의 조화평균으로 정밀도와 재현율이 비슷하면 해당 점수가 높다.

모델에 의해 분류된 진료과목 별로 평가 척도의 평균을 산출하면 정확도는 0.73, 평균 정밀도는 0.78, 평균 재현율은 0.70, 평균 F1-score는 0.73이 나왔다. 개별 진료과목에 대한 정밀도와 재현율, F1-score를 정리하면 <Table 2>와 같다.

<Table 2> 분류모델 결과 표

진료과목	precision	recall	F1-score
신경과	0.71	0.83	0.77
정형외과	0.68	0.74	0.71
외과	0.67	0.69	0.68
내과	0.66	0.52	0.58
이비인후과	0.77	0.68	0.73
산부인과	0.90	0.81	0.85
비뇨기과	0.86	0.72	0.78
안과	0.97	0.63	0.77

F1-score를 보았을 때 산부인과가 0.85로 예측 성능이 가장 높고 비뇨기과가 0.78로 예측 성능이 꽤 높게 나왔음을 볼 수 있다. 이는 산부인과와 비뇨기과의 증상 텍스트에서 중요하거나 자주 나타나는 단어들이 다른 진료 분야에 비해 상대적으로 차별적인 특징을 지니고 있기 때문인 것으로 보인다.

신경과 역시 0.77으로 좋은 예측 성능을 가지고 있으나 <Figure 7>에서 볼 수 있듯이 데이터의 약 1/10을 정형외과로 잘못 분류하고 있다. 이는 신경과 환자들이 호소하는 증상이 정형외과에서 증상을 호소할 때 사용하는 단어 표현을 일부 공유하고 있기 때문일 것으로 판단된다. 데이터가 가장 적었던 안과의 경우 0.77의 F1-score를 기록한다. 이는 <figure 7>에서 볼 수 있듯이 모델이 안과라고 예측한 31개의 데이터는 1개를 제외하고 모두 잘 예측하여 0.97의 가장 높은 정밀도를 가지지만, 실제 49개의 안과 데이터에서 모델이 31개의 데이터만 잘 예측하여 0.63의 재현율을 가졌기 때문이다. 가장 낮은

Confusion Matrix

	신경과 (5.10)	정신외과 (3.58)	외과 (2.65)	내과 (2.33)	이비인후과 (2.25)	심부전과 (1.60)	폐보기관 (1.14)	안과 (0.89)
신경과 (5.10)	425	48	7	10	18	0	2	0
정신외과 (3.58)	60	265	18	12	2	1	0	0
외과 (2.65)	9	41	180	14	14	1	1	0
내과 (2.33)	53	22	24	121	7	5	1	0
이비인후과 (2.25)	35	4	19	12	154	0	0	1
심부전과 (1.60)	0	5	11	5	0	130	9	0
폐보기관 (1.14)	5	4	7	7	1	8	82	0
안과 (0.89)	11	0	3	1	3	0	0	31

<Figure 7> 테스트 데이터 예측에 대한 혼동행렬

F1-score를 기록한 내과의 경우, <Table 1>과 <Figure 7>에서 볼 수 있듯이 내과로 내원해야 하는 환자들이 호소하는 증상이 나머지 7개 진료과목 중 특히 신경과로 내원해야 하는 환자들의 증상과 비슷한 용어를 사용하기 때문인 것으로 보인다.

V. 결론

본 연구에서는 환자들이 호소한 증상 텍스트 데이터를 대상으로 한 진료과목 분류 모델을 구축하였다. 분류 모델을 구축하기 전, 환자들이 네이버 지식인에 주관적으로 서술한 증상 텍스트 데이터의 타당성을 갖추기 위해 진료과목 별 주관적 증상과 진료과목 별 자주 진찰되는 질환의 주요 증상이 정리되어 있는 객관적 증상 문서 간 유사도 분석을 수행하였다. 그 후 tf-idf 벡

터화한 주관적 증상 데이터에서 진료과목 별 단어 중요도를 살펴봄으로써 진료과목 별로 증상 텍스트의 분류가 명확한지 확인해보고자 했다. 해당 데이터를 활용하여 Ridge분류기를 통한 다중 클래스 분류 모델을 학습하여 8개 진료과목에 대해 증상 텍스트를 분류하도록 하였다.

코사인 유사도 분석 결과, 8개 진료과목의 주관적 증상 문서들 모두 동일한 진료과목에 대한 객관적 증상 문서와 상위 3위 이내의 유사성을 보였으므로 주관적 증상을 분류 모델의 학습데이터로 사용하기 위한 타당성을 확보할 수 있었다. tf-idf 벡터에 따른 단어 중요도를 진료과목 별로 상위 20개로 나열해 확인해 본 결과, 진료과목 간 몇몇의 증상 표현이 겹치는 단어들을 제외하고는 대체적으로 진료과목 별 증상의 특징들을 관찰할 수 있었다. 해당 데이터를 정규화한 릿지 분류기로 분류한 결과 평균적으로

0.73이상의 F1-score와 정확도를 확보하였고 진료과목 별 데이터의 특성에 따라 최대 0.85, 최소 0.58의 F1-score를 기록함을 볼 수 있었다.

본 연구의 의의와 한계는 다음과 같다. 첫째로 객관적 증상과 주관적 증상 문서 간 유사도 분석을 통해 개인적 편차와 주관성이 개입될 수 있는 환자의 증상 텍스트가 진료과목 별 전형적으로 보고되는 증상들과 상대적으로 유사성을 보인다는 것을 확인해 보았다는 데 의의가 있다. 그러나 유사도 분석 시 객관적 증상과 주관적 증상 간 의료 용어 표준화가 이뤄지지 않았기 때문에 용어의 의미까지 고려한 유사도 분석을 수행할 수 없었다는 한계가 있다. 만일 용어 표준화 이슈를 해결한다면 더욱 정확한 유사도 수치를 산출할 수 있을 것으로 기대된다. 둘째로 환자가 직접 서술한 증상 텍스트 데이터를 대상으로 진료과목 분류모델이 의미 있는 성능을 기록할 수 있음을 확인했다. 추후 학습가능한 양질의 데이터의 양을 늘리고 데이터 불균형 문제를 해결한다면 더욱 좋은 성능의 모델을 확보할 수 있을 것으로 생각된다. 마지막으로 상용화 가능한 데이터를 확보할 수 있다면, 의료 소비자가 자신의 증상을 입력하면 그에 적합한 진료과목을 추천해주는 서비스의 가능성을 엿볼 수 있었다. 여기에서 더 나아가 사용자의 위치 정보를 받아 추천된 진료과목을 다루는 주변 병원을 권유해주는 알고리즘을 추가적으로 적용할 수 있다면 의료 소비자들의 판단에 도움을 주는 더욱 유용한 모델로 거듭날 수 있을 것으로 기대된다.

참 고 문 헌

- [1] 김윤영 외 2명, “성별에 따른 강박증상의 위험요인으로서 주관적 건강상태”, 보건의료산업학회, 보건의료산업학회지, 제13권, 제4호, pp243-252, 2019.
- [2] 서민환 외 3명, “건성안증후군 환자의 주관적 증상 중증도와 관련 있는 객관적 지표들”, 대한안과학회, 대한안과학회지, 제58권, 제3호, pp259-267, 2017.
- [3] 서울응급의료정보센터, “1339 질환별 상담자료집”, 질병관리청 국가건강정보포털, 2007, (2021.04.10.).
- [4] 유중하, “두통의 증상 및 치료”, 대한 스트레스학회, 스트레스연구, 제3권, 제1호, pp. 1-10, 1995.
- [5] 황상훈, 김도현, “한국어 기술문서 분석을 위한 BERT 기반의 분류모델”, 한국전자거래학회지, 제25권, 제1호, pp.203-214, 2020.
- [6] Afzal, Naveed et al. “Mining peripheral arterial disease cases from narrative clinical notes using natural language processing.” Journal of vascular surgery vol. 65,6, 2017.
- [7] Aurelien Geron, “Hands-On Machine Learning with Scikit-Learn, Keras&TensorFlow”, OREILLY,2018.
- [8] Bo Wang et al., AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system, Applied Soft Computing, Volume 98,2021.
- [9] Castro, Victor M et al. “Large-scale identification of patients with cerebral aneurysms using natural language processing.” Neurology vol. 88,2, pp 164-168, 2017.
- [10] F. Li, D. Cheng and M. Liu, “Alzheimer’s disease classification based on combination of multi-model convolutional networks,” 2017 IEEE International Conference on Imaging Systems and Techniques (IST), 2017, pp. 1-5.
- [11] Hobson Lane, Cole Howard, Hannes Hapke, “Natural Language Processing in Action”, MANNING, 2019.

- [12] Minghao Piao, 변정용, “SVM을 이용한 건강검진정보 기반 진료과목 예측”, 정보처리학회논문지, 소프트웨어 및 데이터 공학 제6권, 제6호, pp303-308, 2017.
- [13] Rajkomar, Alvin et al. “Scalable and accurate deep learning with electronic health records.” NPJ Digital Medicine 1, 2018.
- [14] Sandeep Tata, Jignesh M.Patel, “Estimating the Selectivity of tf-idf based Cosine Similarity Predicates”, SIGMOD Record, Vol.36, No.2, 2007.
- [15] Yang Eun Joo et al., “Symptom Perception and Functioning in Patients with Advanced Cancer”, PLOS ONE, 2021, <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0245987>

저자 소개



이 서 희(Seohee Lee)

- 2017년~현재 : 아주대학교 인문대학 영어영문학과, e-비즈니스학과 재학
- 관심분야 : 텍스트 마이닝, 자연어 처리, 빅데이터 분석, 머신러닝, 딥러닝



강 주 영(Juyoung Kang)

- 1995년 : 포항공과대학교 컴퓨터공학과 (공학사)
- 1997년 : 서울대학교 컴퓨터공학 (석사)
- 2005년 : 한국과학기술원 경영공학 (박사)
- 2005년~현재 : 아주대학교 경영대학 e-비즈니스학과 교수
- 관심분야 : 텍스트 마이닝, 빅데이터 분석, 클라우드 컴퓨팅, 지능형 전자상거래