TABLE OF CONTENTS

LISTS OF FIGURE AND TABLE

ABSTRACT

The Content Analysis System on Cancer-related Medical Articles

국내 암 발생자 수 및 사망자 수가 꾸준히 증가하는 가운데, 환자들은 자신의 치료에 도움이 되는 정보를 찾기 위해 여러 매체들을 통해 정보를 수용한다. 이와 같은 환자들의 니즈에 맞춰 언론 매체들은 수많은 암 관련 의료 기사를 보도하고 있고, 그 영향력이 날로 증가하고 있다. 하지만, 일부 의료기사들은 과학적 근거가 불충분하거나 의약품의 효능을 과장하였으며, 특정 병원 및 의약품의 홍보성 기사들 또한 증가하고 있다. 이러한 기사들은 환자들에게 잘못된 정보를 제공하여 악영향을 일으키게 된다.

올바른 정보를 제공하기 위해 각 병원에서는 의료진이 직접 기사의 적절성을 판단하여 기사를 분류한 후 환자들에게 정보를 제공하지만, 해당 병원에 내방하거나 각 병원 홈페이지에 접속하여 확인해야 하는 등 일반 환자들이 적절성이 검증된 정보를 얻는데에는 어려움이 있다. 또한, 의료진의 주관적 판단 하에 분류되어 적절성을 평가하는 기준이 모호하고, 기사를 읽고 적절성 판단 후 환자에게 제공하기까지의 많은 시간이 소요되어 환자들이 적절한 정보를 빠르게 제공받는데 한계가 있다.

본 연구는 2014년 1월 1일부터 12월 31일까지 한 해 동안 보도된 암 관련 기사 데이터와 의료진에 의해 적절성 판단 후 게재된 서울 모 종합병원에서 제공하는 기사와의 관계를 분석하여, 의료기사 적절성에 영향을 미치는 요인을 찾고 이를 활용해 적절성을 예측하여 기사를 분류하는 시스템을 제안하고자 한다.

KEYWORDS : Text Mining, Cancer-related Medical Articles, Decision Tree, Classification

Ⅰ. Introduction

1. Background and necessity of research

국가 암 정보 센터에서 제공하는 우리나라의 연도별 암 발생률 조사 결과 2012년 10만 명 당 319.5명의 암 발생자 수를 보이며 연평균 3.5%의 증가율로 암 발생률이 증가하는 것을 확인할 수 있다(Jung, 2012). 암 발생률이 증가하면서 환자들은 건강 정보에 대한 관심이 증가하였고, 자신의 치료에 도움이 되는 정보를 찾기 위해 다양한 매체들을 찾으며 정보를 수용하게 되었다. 언론 매체들은 환자들의 니즈에 맞추어 암 관련 의료 기사를 제공하면서 기사의 수와 영향력이 증가하고 있다. 하지만, 본래의 목적과는 다르게 과학적 근거가 불충분하거나 치료 방법 및 의약품의 효능을 과장, 그리고 특정 병원 및 의약품의 홍보를 위한 홍보성 기사들 또한 증가하면서 역기능이 발생하고 있다.

정보를 수용하는 환자들을 위해 일부 병원에서는 의료진이 기사의 적절성을 판단하여 검증된 정보를 환자들에게 제공하고 있지만, 해당 병원에 내방하거나 홈페이지에 접속하여 확인해야 하는 등 일반 환자들이 검증된 정보를 얻는데 큰 어려움이 있다. 또한 의료 기사 보도의 적절성을 판단하는 기준에 대해 연구들이 진행되었지만, 매일 새로운 기사가 보도되는 가운데 환자들에게 검증된 정보를 빠르게 제공할 수 있는 해결책을 제시하지 못하였다. 본 연구는 기존의 적

절성을 평가하는 방법 중 글자 수를 기준으로 판단하였던 방법에서 착안하여 기계학습 방법을 활용하여 전체 기사와 적절성 분석 후 분류된 기사와의 관계를 찾고, 이를 통해 환자에게 적합한 기사를 분류하는 시스템을 제안하고자 한다.

2. Details and method of research

기존의 적절성 평가는 의료진이 직접 의료 기사를 확인하여 적절성을 평가하는 방법으로 진행되었다. 이와 같은 방법은 의료진이 기사를 평가하고 게재하는 과정에서 환자들에게 정보를 신속하게 전달하지 못하는 문제점을 가지고 있다. 본 연구는 머신러닝 기법을 통해 전체 의료기사와 적절성 평가 후 게재된 의료기사의 관계를 분석하여 적절성 판단의 기준을 찾고, 이를 통해 기사의 적절성을 예측하는 시스템을 제안한다.

본 연구는 데이터 수집, 데이터 전처리 과정, 변수 설정, 전체 의료기사와 적절성 평가 후 게재된 데이터와의 관계 분석, 5-fold cross validation을 통한 알고리즘의 검증, 마지막으로 제안하는 알고리즘을 통한 기사의 적절성 분류와 같은 순서로 진행되었다. 전체 의료 기사는 2014년 1월 1일부터 12월 31일까지 한 해 동안 온라인을 통해 보도된 의료 기사들을 수집하였고, 분류의 기준이 되는 기사는 서울 모 종합병원 의료진에 의해 적절성 평가 후 병원 홈페이지에 게재된 기사를 사용하였다. 실험을 위한 데이터로는 2015년 온라인을 통해 보도된 의료 기사들을 사용하였다.

본 연구의 research question은 다음과 같다.

(1) 기사 내 단어의 양이 의료기사의 적절성 판단에 영향을 미치는가?

(2) 기사 내 빈도수가 높은 단어가 의료기사의 적절성 판단에 영향을 미치는가?

(3) 기사의 제목에 등장하는 단어와 본문 내 단어와의 연관성이 의료기사의 적절성 판단에 영향을 미치는가?

본 논문의 구성은 다음과 같다. 2장에서는 의료기사가 환자들에게 미치는 영향과 기존의 적절성 판단에 대한 연구에 대해 소개하고, 연구에 사용되는 텍스트 마이닝, Decision Tree, k-fold cross validation에 대해 살펴본다. 3장에서는 본 논문에서 제안하는 시스템 구현 절차에 대해 소개한다. 4장에서는 제안하는 시스템을 통해 예측 되어진 기사를 사용하여 의료진을 대상으로 실험한 결과를 소개한다. 5장에서는 연구 결과를 바탕으로 결론 및 한계점, 그리고 향후 연구 방향을 제시한다.

Ⅱ. Related Work

1. Problems of Cancer-related Medical articles

고령화 사회의 진행 및 국내 암 환자 수의 증가로 인해 건강 정보에 대한 국민들의 관심이 크게 증가하고 있다. 특히 암 환자들에게 미디어에서 제공하는 의학 기사들은 새로운 치료 방법이나 암 예방 방법에 대한 정보를 얻는데 주요한 방법이 되었다(Rodolf, 2004). 건강보험공단의 설문조사에 따르면 국민들이 건강에 대한 정보를 접하는 방법에 있어서 인터넷이 75%로 의료인 7.1%보다 크게 앞서고 있다(연합뉴스 2010.12.10). 이러한 국민들의 니즈에 맞춰 언론사에서는 온라인을 통해 누구나 쉽게 건강 관련 의료 기사들을 보도하고 있고, 사람들은 대부분의 건강 정보를 온라인을 통해 얻고 있다. 하지만 의료 기사의 수와 그 영향력이 날로 증가하는 가운데 검증되지 않은 의료 정보들 또한 증가하고 있다(류시원, 2003). 한 연구결과에 따르면 방송에 노출된 의학 정보 중 40%가 부정확하거나 오해를 유발 할 수 있는 정보로, 이는 기자와 의사와의 관계, 광고게재 등의 변수에 의한 이해관계에 따른 기사 게재가 많음을 보여주고 있다. 이러한 정보들은 과학적 근거가 불충분하거나 치료 방법 및 의약품의 효능을 과장, 그리고 특정 병원 및 의약품의 홍보를 위한 홍보성 기사들로 환자들에게 역기능을 초래할 수 있다(김길원, 2008).

2. 암 관련 보도 기사량(글자 수)에 따른 적절성 분석

암 관련 보도의 역기능이 증가하면서, 의료 기사의 적절성을 판단하는 기준을 찾기 위한 연구들이 진행되었다. 박정의(2002)는 의료 기사의 기사 유형, 기사량, 기사 작성자, 정보의 출처에 따른 의학적 건전성, 필수적 정보의 누락, 기사의 과장을 조사하여 암 관련 보도의 적절성 판단에 영향을 미치는 요소를 찾기 위한 연구를 진행하였다. 총 752개의 의료 기사를 수집하여 기사량과 의료 기사 적절성 사이의 관계 분석한 결과, x^2=22.581, p<0.001의 통계적으로 유의한 결과를 보였다. 기사량이 증가함에 따라 의학적 건전성이 증가하며, 필수적 정보의 누락 및 기사 정보의 과장은 줄어드는 등 기사량이 의료 기사의 적절성에 영향을 미친다는 것을 알 수 있었다(박정의, 2002). 의료 기사는 주제가 방대하여 공통된 분류 기준을 정하는 것이 어렵다는 점에서 기사량을 통한 기사의 적절성 판단은 의미 있는 결과이다. 하지만 전체 의료 기사와, 적절성 평가 후 보도된 기사와의 관계를 글자수를 이용해 비교해 본 결과, 그림1과 같이 다수의 outlier가 나타나는 문제가 발생하였다. 본 연구는 이와 같은 문제를 해결하기 위해 글자 수 외에 단어의 양 및 특정 단어의 빈도수 그리고 단어 간의 연관성 분석을 더하여 기사의 적절성을 판단하는 방법을 제안한다.

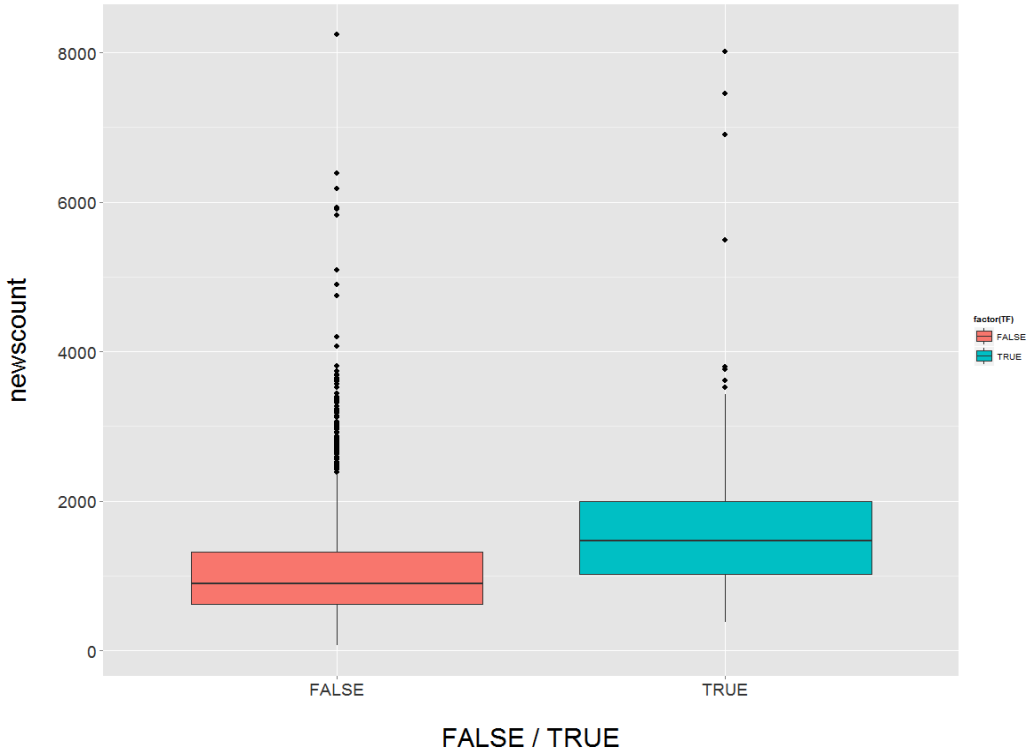

Figure 1. 기사량(글자수)에 따른 True, False Boxplot

3. Text Mining Process

텍스트 마이닝은 텍스트로 구성된 다양한 데이터로부터 유용한 정보와 지식을 추출하는 과정이다. 텍스트 정보의 양이 증가하면서 효율적인 정보의 추출 및 관리를 위해 사용되고 있다. 텍스트 마이닝은 문서의 정리 및 자동분류, 소비자 니즈 분석 등 다양한 분야에서 사용되고 있다. 분류를 위한 텍스트 마이닝의 단계는 크게 데이터 수집, 전처리과정, 문서 분류 모델 구축, 분류 프로세스로 나눌 수 있다.

A. Preprocessing

Preprocessing is a process for creating the index file. It makes easier to handle the text data. There are three ways to make an index-file, they are FRB(Frequency-Based), IDF(Inverse Document Frequency), LSI(Latent Semantic Indexing).

FRB(Frequency-Based) is a simple way to make index-file. This method identifies important words using the way to give higher weight to the word. IDF(Inverse Document Frequency) is a method complements the FRB. It is not only finds important words but comprise words distinguished from the other document. LSI(Latent Semantic Indexing) is a method for making possible to search the document by topic or concept.

이 외에도 전처리과정은 다양한 방법을 통해 데이터를 사용하기 쉽게 가공한다. 본 연구에서는 Extract Noun, Morphological segmentation 등을 위한 Natural Language Procesing을 통해 기사 내의 단어들을 추출하는 과정을 시행하였다. 또한, 홍보성 기사를 제거하기 위해 특정 병원 및 상품의 이름을 제거하는 과정을 시행하였다.

B. Association rule algorithm

연관성 분석은 데이터 내에 각 항목간의 상관관계를 찾아내는 과정이다. 연관성을 평가하는 기준은 지지도(Support), 신뢰도(Confidence), 향상도(Lift)가 있다. Support는 전체 데이터 중 조건부와 결과부 모두를 만족하는 비율을 의미한다. Confidence는 조건부를 만족하는 데이터 내에서 조건부와 결과부를 만족하는 데이터의 비율이다. Lift는 조건부 없이 결과부를 만족할 확률 대비 조건부가 주

어졌을 때 결과부를 만족할 확률이다(Son, 2014). 연관성 분석은 Support, Confidence, Lift 순서로 시행된다. 각 기준에 대한 공식은 표1과 같다.

지표	수식
Support	$P(A \cap B)$
Confidence	$P(B\|A) = \frac{P(A \cap B)}{P(A)}$
Lift	$\frac{P(B\|A)}{P(B)} = \frac{P(A \cap B)}{P(A)P(B)}$

Table 1. Formula of Association rule algorithm

본 연구에서는 연관성 분석을 사용하여 기사 내에 등장하는 암이라는 단어가 포함된 문장 내 단어들과, 제목 문장 내 단어들간의 관계를 분석하였고, 이를 바탕으로 각 기사의 적절성을 판단하였다. 만약, 암이라는 단어가 기사 내에 등장하지 않을 경우, 기사 내 최대 빈도수의 단어를 기준으로 정하여 분석을 시행하였다.

C. Decision Tree.

결정 트리 학습법은 분류를 위한 학습 방법 중 하나이다. Decision tree는 데

이터의 특징을 찾아내어 각 속성의 조합으로 분류모형을 나타낸다. 목표 값에 대해 관련이 높은 속성들이 위에서부터 계층적 구조로 나타나기 때문에 모형결과를 이해하기 쉽다는 장점을 가지고 있다. Decision tree를 활용한 분석 과정은 데이터의 분류를 위한 feature vector 설정, decision tree 가지치기, 모형 검증, 예측 순으로 이루어진다.

그림 2는 decision tree의 예시이다. 입력된 데이터는 A, B, C, B′이라는 feature vector의 속성에 부합하는지에 따라 왼쪽, 혹은 오른쪽 가지로 이동한다. 그 결과 최종적으로 b, b″, b‴, c′, c″ 이라는 값으로 분류되게 된다.

본 논문에서는 의료기사라는 점에서 환자에게 잘못된 정보를 전달할 경우 환자의 건강에 치명적인 문제가 발생한다고 판단하여, 환자에게 제공되는 정보가 옳다는 귀무가설 하에 귀무가설이 옳음에도 그를 기각하는 제 1종 오류를 높이는 방법을 선택하였다. 따라서 본 알고리즘에서는 기사가 적절하지 않다고 판단하는 FALSE의 비율이 높아지도록 설계하였다.

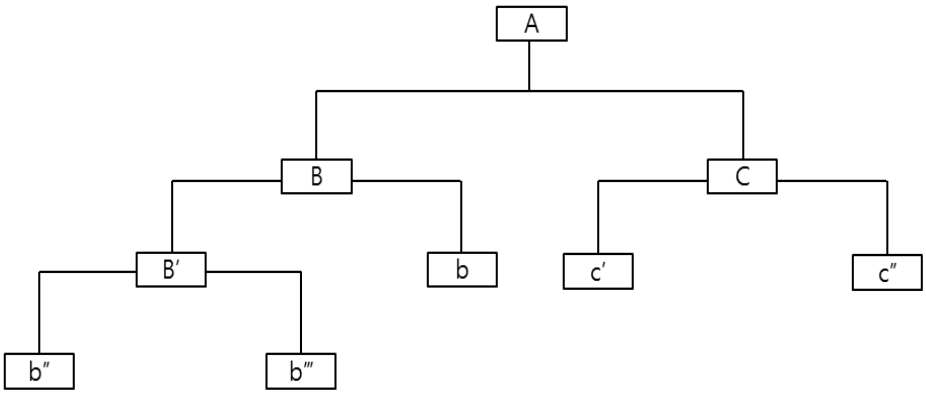

Figure 2. Decision tree model

D. K-fold Cross Validation.

k-묶음 교차 검증법은 분류모델의 검증을 위한 분석 방법이다(김형도,2014). 대량의 데이터를 수집할 수 없는 상황에서 분류기의 적합도를 평가하는 것은 정확도가 떨어지고, Overfitting의 문제가 발생할 수 있다. 이러한 문제점을 해결하기 위해 k-fold cross validation은 데이터를 k개의 집합으로 분할한 후, 하나의 집합을 test set으로, k-1개의 집합을 training set으로 나눈다. 각 집합들은 k번 바뀌면서 학습과 검증과정을 거치며 k개의 다른 모형을 구축하게 되고 각각의 정확도를 비교하며 모델을 검증하게 된다(Braga,2014). 본 연구에서는 그림3과 같은 5-fold cross validation을 사용하였다. 먼저 데이터를 5개의 집합으로 나눈 후, test set과 training set을 5회 바꾸며 분류 모형을 검증하였다.

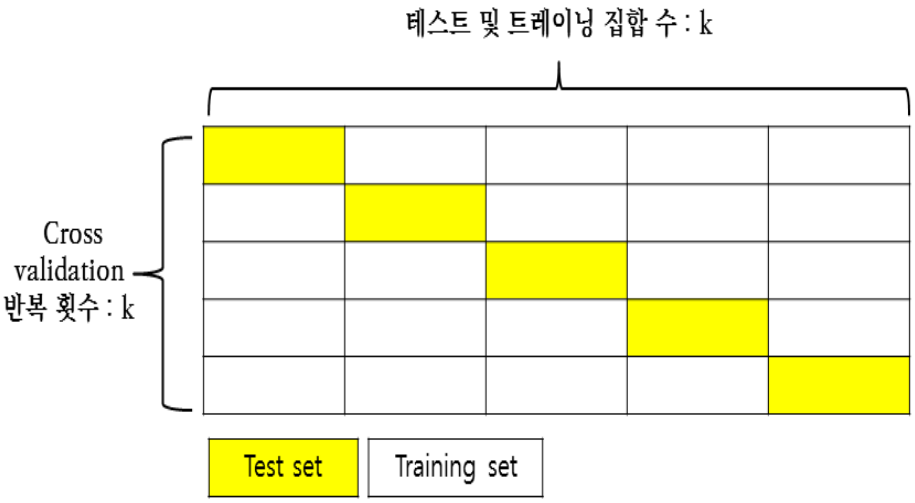

Figure 3. k-fold cross validation

Ⅲ. Research Method

1. Approach

본 연구는 통계 프로그램인 R을 사용하여 다음과 같은 프로세스로 진행되었다. 첫 번째는, 분류 기준을 찾기 위해 2014년 한 해 동안 보도된 암 관련 의학 기사와, 의료진에 의해 적절하다고 판단된 기사, 제안하는 최종 분류 모델 검증에 사용 될 2015년에 보도된 암 관련 의료 기사 등 데이터의 수집 과정이다. 두 번째는, 데이터 모델링을 효율적으로 진행하기 위해 데이터 전처리 과정을 진행한다. 이 과정에서는 명사 추출, 형태소 분석 등의 자연어 처리 과정과, 특정 병원 및 의약품의 이름, 제약회사의 이름 등 병원 및 제품 홍보를 위해 쓰여진 적절하지 못한 기사를 분류하기 위해 관련 단어들을 제거하였다. 세 번째는, 분류 기준이 되는 feature vector를 설정하였고, Decision tree 기법을 통해 분류 모델링을 진행하였다. 이 과정을 통해 어떠한 feature vector가 적절성 판단에 영향을 미치는지 알아보았고 k-fold cross validation을 통해 분류 모델이 적합한지 판단하는 과정을 거쳤다. 마지막으로 네 번째는, 분류 모델링에 사용되지 않은 2015년 기사를 입력하여, 제안하는 분류 기준을 통해 분류 되어진 기사를 의료진에 의해 적절성 평가를 받는 과정을 거쳤다. 본 연구의 과정은 그림 4와 같다.

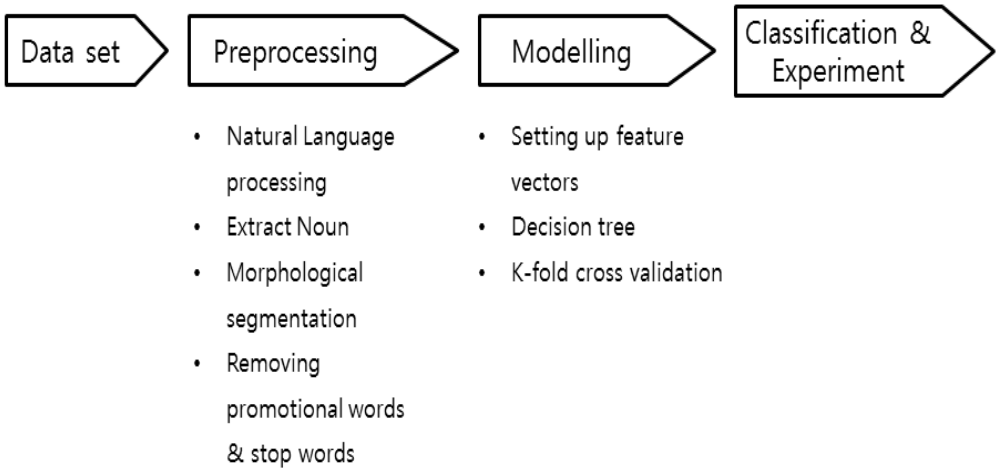

Figure 4. Research process

2. Data Source

본 연구는 2014년 한 해 동안 온라인에 보도된 암 관련 의료 기사와 서울 모 종합병원에서 의료진의 적절성 평가 후 환자들에게 제공되는 기사를 사용하여 분석되었다. 현재 병원과 언론사에서 의학기사의 적절성을 판단하는 명확하고 공통된 기준이 부재하여 서울 모 종합병원 의료진이 판단한 기사들을 적절한 기사의 기준으로 사용하였다. 적절성 평가 기준은 기사의 양이 일정 수준 이상인가, 제목과 기사 본문의 내용이 일치하는가, 광고, 홍보성 내용이 포함되지 않았는가, 기사 내용이 일관적인가, 환자에게 적절한 기사인가로 구성되었다.

2014년 전체 암 관련 기사는 유방암, 폐암, 간암 등의 암 종류가 포함된 기사이며 총 3832건의 기사를 수집하였다. 이 중 서울 모 종합병원 의료진에 의

해 적절하다고 평가된 기사는 총 1346건 이었다. 수집된 기사는 그림5와 같이 날짜, 제목, 기사 내용으로 정리하였다.

date	title	news
2014.06.01_01	항암제 '렌바티닙', 방사성요오드치료	에자이의 항암제 '렌바티닙
2014.06.01_02	아스트라제네카 난소암 치료 신약 효	암 세포가 필요로 하는 신
2014.06.01_03	울산대병원, 혈액형 불일치 간이식 수	지역 최초 시행 울산대학교
2014.06.01_04	릴리 위암 치료제 '폐암' 환자 수명도	릴리사의 함암 치료제가 폐
2014.06.01_05	존슨앤존스 백혈병 치료 신약 말기 임	파마싸이클릭스(Pharmacy
2014.06.01_06	전이성 척추종양 환자 생존율 높였다	건양대병원 이창현 교수팀
2014.06.01_07	소식(小食)하면 몸 속 암세포 줄어든	평소 적은 칼로리를 섭취하
2014.06.02_01	자궁경부암에 대한 면역 요법 성공 보	인간유두종 바이러스(HPV
2014.06.02_02	사노피-서울아산, 간암환자 공동연구	사노피와 서울아산병원이
2014.06.02_03	식이요법과 운동 '병행'해야 유방암 (	폐경이 된 여성의 경우, 같
2014.06.02_04	원자력의학원, 유방암환우회 바자회	유방암환자 식사 관리 강연
2014.06.02_05	흡연이 사망율 50% 높여, 男폐암환자	담배를 피우면 한국인 등
2014.06.02 06	재발 난소암 화학요법 대안 찾았다	<U+00A0>2종 약물조합을

Figure 5. Cancer-related Medical articles data in 2014

본 연구에서 제안하는 최종 모델의 검증에 사용되는 기사는 모델링 과정에 사용되지 않은 2015년 온라인에 보도된 암 관련 의료 기사로, 총 109건의 기사를 수집하였다.

3. Preprocessing

데이터 모델링을 효율적으로 진행하기 위해 데이터 전처리 과정을 진행하였다. 이 과정에서는 명사 추출, 형태소 분석 등의 자연어 처리 과정과, 특정 병원 및 의약품의 이름, 제약회사의 이름 등 병원 및 제품 홍보를 위해 쓰여진 적절하지 못한 기사를 분류하기 위해 관련 단어들을 제거하였다. 그림6은 어미, 조사 등을 제거하고 명사와 형용사만을 추출한 결과이다. 이 결과를 바탕으로 의료진과의 검증과정을 통해 부적절한 단어와 특정 회사명 등을 제거하였다.

fileindex	word
2014-01-0	콜레스테롤 제거 혈관 깨끗 보호 새해 연초 각종 모임 식이요법 성인 환자 고량진미 유혹 평생 음식 조절 성인 환자 경우 증상 악화 음
2014-01-0	임신 유방 경우 임신 유방 유방 사춘기 이후 평생 수유 준비 기관 임신 유방 조직 평소 정도 혈관 증식 임신 수유 여성 들이 뭉치 이유
2014-01-0	금연 개월 다이어트 고비 새해 신년 계획 금연 어트일 금연 결심 남편 어트를 결심 아내 조심 시기 결심 개월 작심삼일 계획 지속 해서
2014-01-0	비타민 가득 건강학 현미 백미 기피 사람 식이 섬유 보고 진행 당뇨병 치료제 건강 조선일보 영양가 현미 백미 쌀겨 때문 소화 하기 면
2014-01-0	여성 금연 클리닉 프로그램 활성화해 담배 남녀 불문 모두 여성 흡연 남성 부정 시각 이유 생리학적 아기 임신 여성 특성상 흡연 건강
2014-01-0	일사천리 생활 건강 어트 금연 운동 새해 남녀노소 누구 새해 결심 목표 공통점 새해 건강 소망 새해 건강 열풍 명예 건강 전부 건강 우
2014-01-0	암연구 담배 인상 흡연율 담배세 인상 흡연율 연구 결과 영국 암연구 소속 과학자 뉴잉글랜드 의학 저널 제출 논문 세계 담배세 현재
2014-01-0	스트레스 침분비 감소 구내염 원인 과로 스트레스 입안 상처 사람 입속 상처 통칭 구내염이라고 구내염 구강 관련 부위 염증 통칭 입술
2014-01-0	새해벽두 피부 일종 흑색종 증상 모양 비슷 일쑤 정확 진단 치료 목표 피부 흑색종 치료 삼성 서울병원 새해 클리닉 개설 본격 치료 멜
2014-01-0	삼성 서울병원 흑색종 피부 클리닉 개설 삼성 서울병원 대표 피부 흑색종 치료 클리닉 새해 개설 본격 치료 멜라닌 세포 암세포 반점
2014-01-0	청순한 한복 자태 암환자 호중구성발열 치료제 칸시다스주 보험 급여 보건복지부 중증 질환 보장 강화 차원 호중구감소성 발열 칸시다
2014-01-0	고용량 비타민 알츠하이머 진행 지연 고용량 비타민 알츠하이머 증상 진행 지연 연구 결과 발표 미국 미니애폴리스 헬스 케어 시스템
2014-01-0	배달 음식 치킨 칼로리 한국농촌경제연구원이 전국 가구 대상 외식 소비 행태 결과 치킨 배달 음식 치킨 햄버거 핫초코 칼로리 건강 차
2014-01-0	짝수년 출생 건강 검진 받으 청양군보건의료원 무료 성인 검진 국가암검진을 내달 출장 검진 실시 검진 대상 이상 짝수년 출생 의료 급
2014-01-0	흡연자 절반 담배 관련 질병 사망 담배 세금 세계 죽음 연구 결과 영국 암연구소 담배세 젊은이 신규 흡연 기존 흡연자 금연 연구소 선
2014-01-0	가래 질병 연초 건강 검진 사람 변화 건강 검진 유용 가래 손발 질병 예방 헬스 조선 질환 무색투명 가래 급성기관지염 가능성 분홍색
2014-01-0	콜레스테롤 제거 혈관 깨끗 보호 새해 연초 각종 모임 식이요법 성인 환자 고량진미 유혹 평생 음식 조절 성인 환자 경우 증상 악화 음
2014-01-0	운동 장수 근면 성실 기상 갑오년 새해 아침 사람 올해 기운 건강 전문의 생활 습관 질병 건강 미국 의학자 브레슬로와 벨록은 신체 정
2014-01-0	운동 작심삼일 극복 앵커 멘트 새해 건강 여념 모습 새해 목표 마음 문젭니다 금연 대표 작심삼일 새해 목표 연초 담배 판매량 불과 증
2014-01-0	갑상선 생존 치료 결정 갑상선 갑상선호르몬 분비 분비기관 갑상선 호르몬 우리 신진대사 조절 긴장 항진 이완 안정 상태 조성 신체 내
2014-01-0	뇌종양 생존 통계 숫자 진실 폐암 위암 간암 유방암 대장암 각종 생존 추세 가운데 뇌종양 생존 이상 비교 뇌수막종 뇌하수체 선종 신

Figure 6. Result of Preprocessing

4. Modelling

본 연구에서 적절성의 기준이 되는 서울 모 종합병원에서 의료진의 평가 후 게재한 기사의 특징을 찾은 결과, 일정 수준 이상의 글자 수와 단어가 등장했을 때 적절성을 위한 필수 정보들이 존재하였다. 반면에 중복되는 단어를 제거하였을 때 등장한 총 단어의 수가 많을 경우 여러 주제를 다루며 기사 내용이 일관적이지 못하고 필수 정보가 생략되는 것을 알 수 있었다. 또한 기사 내에서 전달하고자 하는 내용이 뚜렷할 경우 상위 빈도수의 단어들이 자주 등장하는 것을 알 수 있었다. 이러한 특징을 바탕으로 의사결정나무에서 데이터를 분류 및 예측하기 위한 특징을 표2와 같이 구성한 후 의사결정나무 기법을 통한 분류 및 예측을 진행하였다.

Feature vector	설명
News count	기사에 등장한 총 글자 수
Volume	기사에 등장한 총 단어의 수
Univolume	기사에 등장하는 단어 중 중복되는 단어를 제거한 총 단어의 수
First	기사 내에 가장 많이 등장한 단어의 수/기사 내 전체 단어의 수 * 100
Second	기사 내에 두 번째로 많이 등장한 단어의 수/기사 내 전체 단어의 수 * 100

Third	기사 내에 세 번째로 많이 등장한 단어의 수/기사 내 전체 단어의 수 * 100
Cancer	기사에서 '암' 이라는 단어의 등장 유무 (등장=1, 등장하지 않음=0)
Lift	기사의 제목에 등장하는 단어와 본문 내 단어와의 연관성 분석 (단어 '암'을 포함한 문장 속 단어와 비교하였으며, 단어 '암'을 포함하지 않을 시에 본문 내 최대 빈도수 단어와 비교)

Table 2. Feature vector

A. Modelling result

그림 9는 본 연구에서 제시하는 알고리즘을 바탕으로 기사의 적절성을 분류한 결과이고, 그림 7은 Decision tree를 결과 모델이다. 3열은 전체 기사 중 기준이 되는 서울 모 종합병원에 게재된 여부에 따라 게재 시 True, 미게재 시 False로 분류하였고, 4열은 본 알고리즘으로 True, False를 분류한 결과이다.

date	title	TF	pre
2014.06.01_01	항암제 '렌바티닙', 방사성요오드치료 저항성	FALSE	FALSE
2014.06.01_02	아스트라제네카 난소암 치료 신약 효과적	FALSE	FALSE
2014.06.01_03	울산대병원, 혈액형 불일치 간이식 수술 성	FALSE	FALSE
2014.06.01_04	릴리 위암 치료제 '폐암' 환자 수명도 연장	FALSE	FALSE
2014.06.01_05	존슨앤존스 백혈병 치료 신약 말기 임상시험	FALSE	FALSE
2014.06.01_06	전이성 척추종양 환자 생존율 높였다	FALSE	FALSE
2014.06.01_07	소식(小食)하면 몸 속 암세포 줄어든다"	FALSE	FALSE
2014.06.02_01	자궁경부암에 대한 면역 요법 성공 보고돼	FALSE	FALSE
2014.06.02_02	사노피-서울아산, 간암환자 공동연구 협약	FALSE	FALSE
2014.06.02_03	식이요법과 운동 '병행'해야 유방암 예방 효	TRUE	FALSE
2014.06.02_04	원자력의학원, 유방암환우회 바자회 열어	FALSE	FALSE
2014.06.02_05	흡연이 사망율 50% 높여, 男폐암환자 60%↑	TRUE	FALSE
2014.06.02_06	재발 난소암 화학요법 대안 찾았다	FALSE	FALSE
2014.06.02_07	유방암 치료제 할라벤, 1일부터 보험급여 적	FALSE	FALSE
2014.06.02_08	자궁경부암 감소 추세?...35세 미만 젊은 환	TRUE	FALSE

Figure 7. Result of classification

Training Set \ Test Set	False	True
False	2475	117
True	575	665

Table 3. Result of Classification

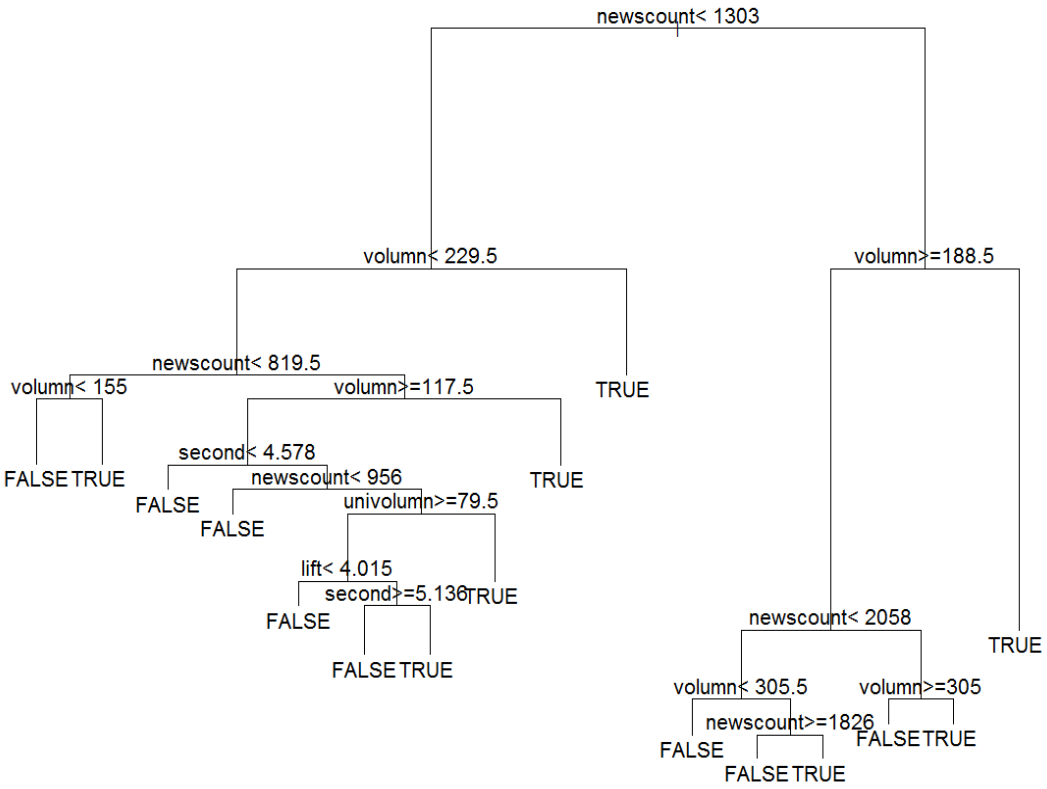

Figure 8. Result of Decision tree

분류 결과 그림 8과 같이 newscount, volume, second, univolume, lift 순으로 기사 분류에 영향을 미치는 것으로 나타났다. 분류 정확도는 82%로 나타났다.

B. 5-fold Cross Validation

본 분류 모형의 검증을 위해 전체 데이터의 80%를 Training data로 구성하고, 20%를 Test data로 구성하여 총 5회의 5-fold Cross Validation을 진행하였다. 각 진행 결과는 다음과 같다.

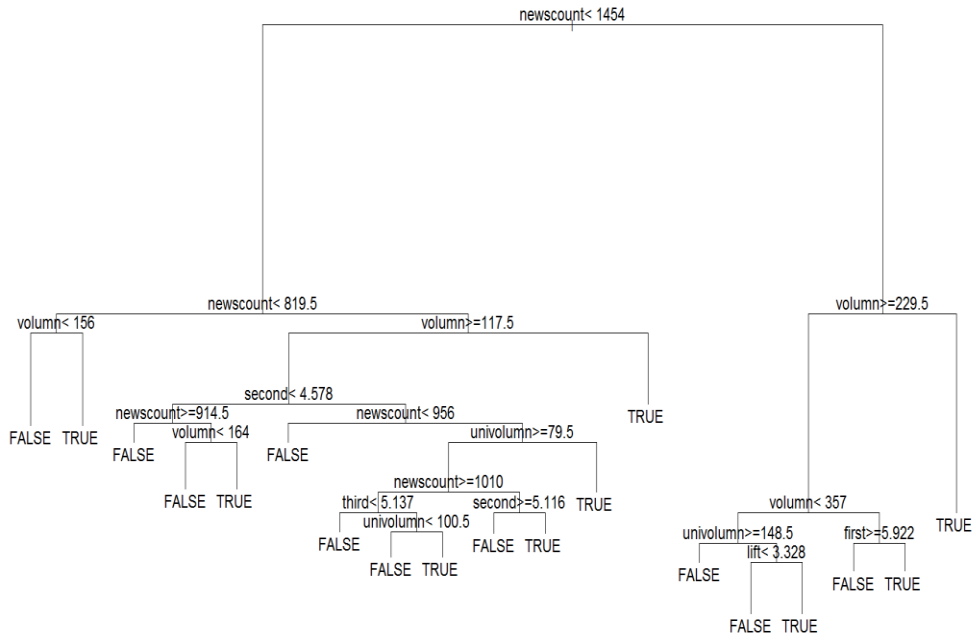

Figure 9. Result of cross validation(1)

Test Set / Training Set	False	True
False	456	142
True	44	125

Table 4. Result of cross validation(1)

첫 번째 결과에서는 그림 9와 같이 newscount, volume, second, univolume, 등의 순서로 기사 분류에 영향을 미치는 것으로 나타났다. 분류 정확도는 76%로 나타났다.

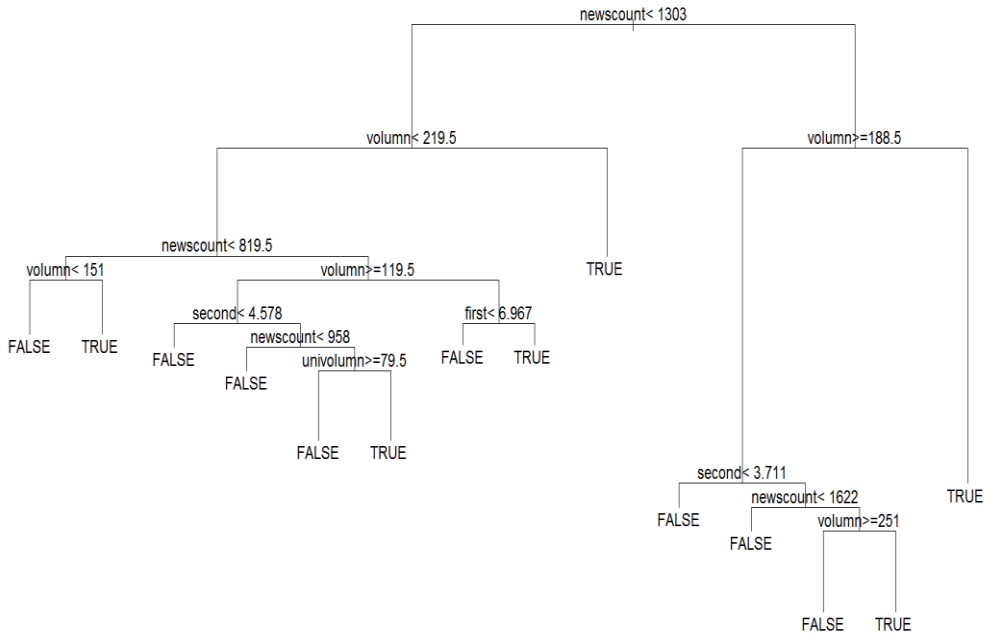

Figure 10. Result of cross validation(2)

Test Set / Training Set	False	True
False	501	127
True	28	111

Table 5. Result of cross validation(2)

두 번째 결과에서는 그림 10과 같이 newscount, volume, second, firist 등의 순서로 기사 분류에 영향을 미치는 것으로 나타났다. 분류 정확도는 80%로 나타났다.

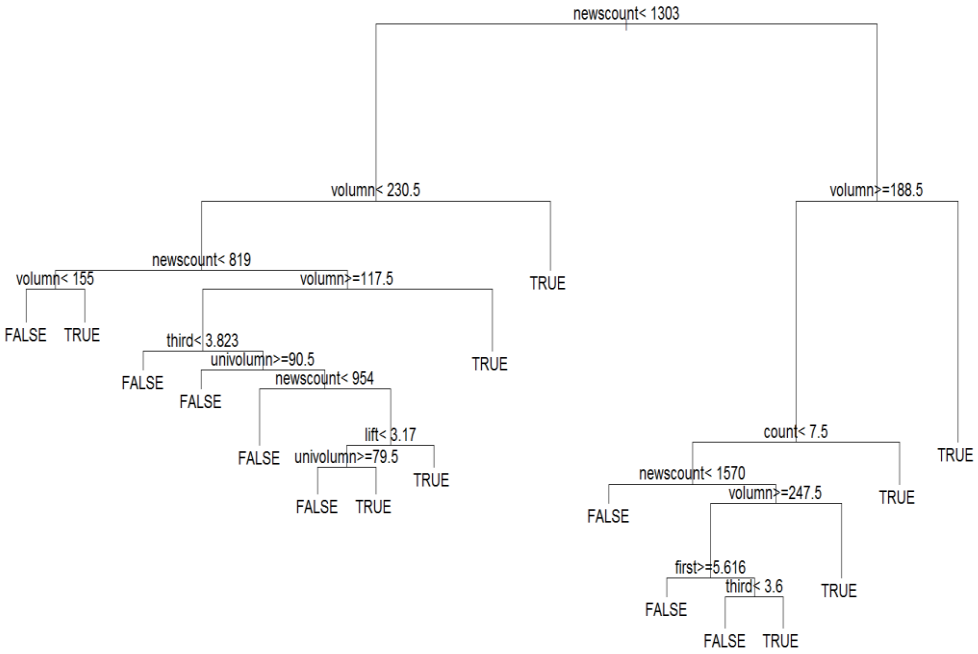

Figure 11. Result of cross validation(3)

Test Set / Training Set	False	True
False	474	107
True	46	139

Table 6. Result of cross validation(3)

세 번째 결과에서는 그림 11과 같이 newscount, volume, third, univolume 등의 순서로 기사 분류에 영향을 미치는 것으로 나타났다. 분류 정확도는 80%로 나타났다.

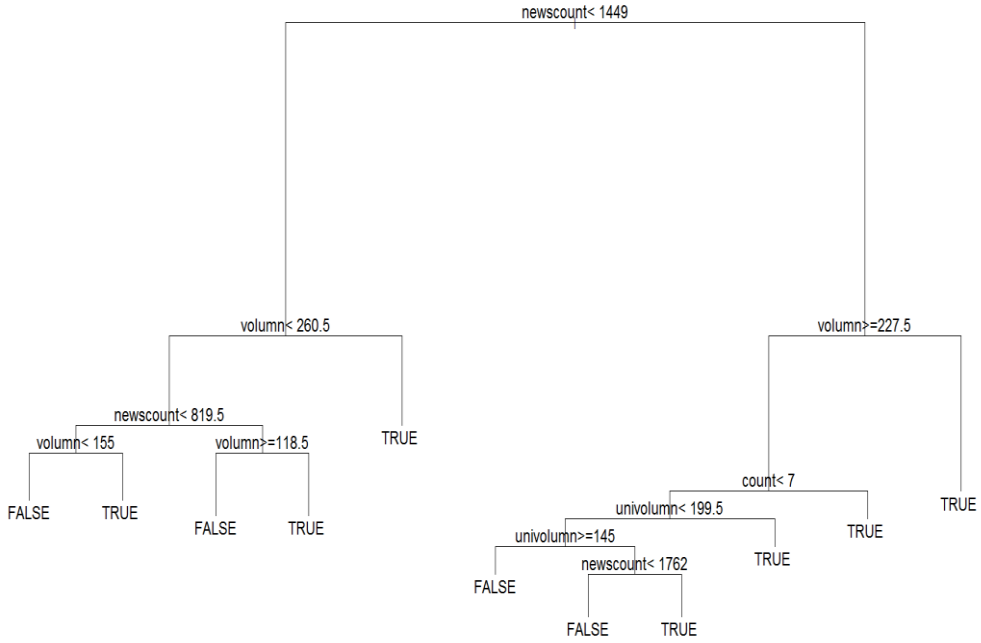

Figure 12. Result of cross validation(4)

Test Set / Training Set	False	True
False	481	135
True	43	107

Table 7. Result of cross validation(4)

네 번째 결과에서는 그림 12와 같이 newscount, volume, count, univolume 등의 순서로 기사 분류에 영향을 미치는 것으로 나타났다. 분류 정확도는 77%로 나타났다.

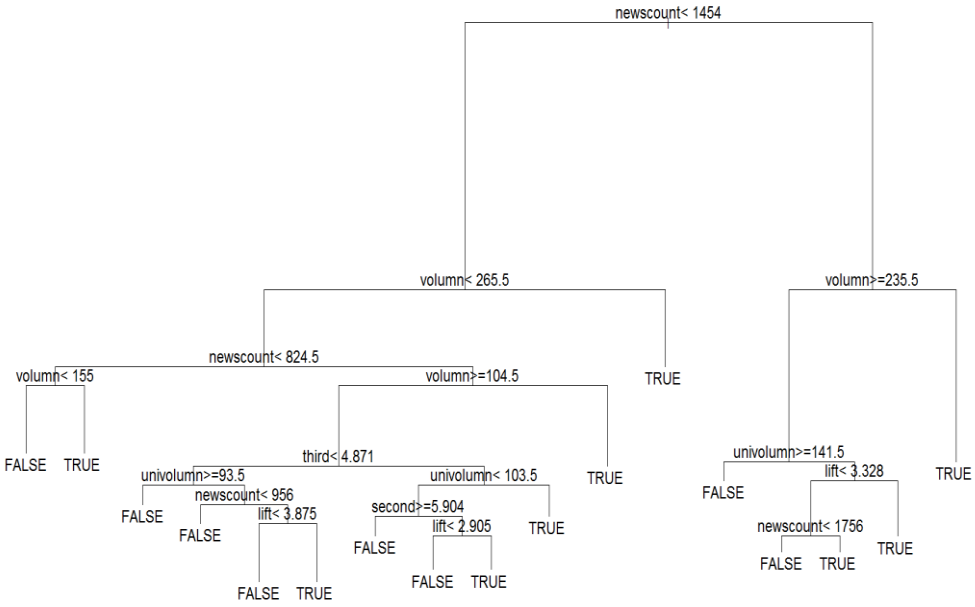

Figure 13. Result of cross validation(5)

Training Set \ Test Set	False	True
False	474	107
True	46	139

Table 8. Result of cross validation(5)

마지막으로, 다섯 번째 결과에서는 그림 13과 같이 newscount, volume, third, univolume, lift 등의 순서로 기사 분류에 영향을 미치는 것으로 나타났다. 분류 정확도는 80%로 나타났다.

5. Classification

본 연구에서 제안하는 모델로부터 분류된 기사의 적절성을 의료진으로부터 검증하기 위해 2015년에 보도된 기사의 적절성 분류를 예측해보는 과정을 진행하였다. 사용된 데이터는 2015년에 보도된 기사 109건으로 그림 14와 같다.

date	title	news
2015.10.26_01	베이컨, 소시지 발암물질, 담배와 동급	WHO, 위협요인 규정 예정<U+0
2015.10.21	내시경 위암 수술, 이산화탄소 이용하면	내시경으로 위암을 치료할 경우
2015.10.28	위암 예방 및 개선 위한 음식 선택법은?	<U+00A0>매일 과중한 업무에
2015.10.07	국내 위암 환자 3명 중 2명은 건강검진	건강검진을 받은 사람이 위암을
2015.09.21	[암과의 동행-질환 통계] 65세 이상 노인	65세 이상 노인에게 많은 암은 위
2015.10.26	별 증상 없는 갑상선암 '조기검진'으로	전업주부 A(48)씨는 종합검진을
2015.09.15	담배, 폐에만 영향준다? NO!...'위암' 예방	가스차고 더부룩 소화불량 착각.
2015.06.30	인스턴트식, 적게 먹어도 위암위험 4.4배	인스턴트식품을 적게라도 섭취하
2015.10.13	위건강 해치는 한국인 식습관... 위, 험한	예로부터 농경사회를 기반으로
2015.10.29	20대 딸이 50대 엄마보다 유방암 위험 2	현재 우리나라의 유방암 증가 추
2015.10.21	미국암학회 새 유방암 권장기준 "45세부	미국암학회(ACS)가 여성들에게
2015.10.01	여성암 1위 유방암, 40세 이상 여성은 2	여성이라면 누구도 안심할 수 없
2015.07.22	유방암 자가진단법, 위 팔이 붓는다면 '의	유방암 자가진단법엔 무엇이 있
2015.07.15	30~40대 여성, 유방암 자가검진법 100	우리나라 여성 30~40의 유방암

Figure 14. Cancer-related Medical articles data in 2015

그림 15는 109개의 기사를 본 논문에서 제안하는 알고리즘에 입력하여 적절성을 예측한 결과이다. 총 10건의 기사가 적절하다고 분류되었다.

date	title	pre
2015.10.26_01	베이컨, 소시지 발암물질, 담배와 동급	FALSE
2015.10.21	내시경 위암 수술, 이산화탄소 이용하면	FALSE
2015.10.28	위암 예방 및 개선 위한 음식 선택법은?	FALSE
2015.10.07	국내 위암 환자 3명 중 2명은 건강검진	FALSE
2015.09.21	[암과의 동행-질환 통계] 65세 이상 노인	FALSE
2015.10.26	별 증상 없는 갑상선암 '조기검진'으로 관	FALSE
2015.09.15	담배, 폐에만 영향준다? NO!...'위암' 예방	FALSE
2015.06.30	인스턴트식, 적게 먹어도 위암위험 4.4배	FALSE
2015.10.13	위건강 해치는 한국인 식습관... 위, 험한	FALSE
2015.10.29	20대 딸이 50대 엄마보다 유방암 위험 2	FALSE
2015.10.21	미국암학회 새 유방암 권장기준 "45세부	TRUE
2015.10.01	여성암 1위 유방암, 40세 이상 여성은 2	FALSE
2015.07.22	유방암 자가진단법, 위 팔이 붓는다면 '의	FALSE
2015.07.15	30~40대 여성, 유방암 자가검진법 100명	TRUE

Figure 15. Result of classification(2)

6. Experiment

본 논문에서 제안하는 알고리즘을 통해 적절하다고 분류된 기사를 의료진에게 검증하는 과정을 실시하였다. 실험인은 서울 모 종합병원 2곳에서 근무 중인 4명으로 구성되었다. 질문의 구성은 다음과 같다.

1) 기사의 양이 적절한가?

2) 제목과 기사의 내용이 관련있는가?

3) 광고, 홍보성 내용이 포함되지 않았는가?

4) 환자에게 적합한 기사인가?

5) 기타 의견

앞선 분류에서 적절하다고 분류된 10가지 기사에 대해 다음과 같은 질문으로 기사의 적절성을 평가한 결과, 6개의 기사가 환자에게 적절하다고 평가되었다. 다음은 환자에게 적합한 기사인가를 종속변수로 설정한 후, 기사의 양, 제목과 기사의 내용 관련성, 광고 홍보성 내용이 포함되었는가의 질문을 독립변수로 설정하여 어떠한 요소가 기사의 적절성에 영향을 미치는지 분석하였다. 그 결과, R^2=0.878, P=0.04로 회귀모형에 적합한 결과가 나왔고, 제목과 기사의 연관성(P=0.007)이 환자에게 적합한 기사인지에 대한 가장 큰 영향을 미치는 것을 알 수 있었다. 또한, 광고, 홍보성 내용의 유무(P=0.139)는 환자에게 적합한 기사인지에 대해 영향을 미치지 않는 것으로 나타났다.

Ⅳ. Conclusion

본 논문은 기사의 적절성을 판단하는 기준 중 글자수만을 활용하던 기존의 연구를 개선하기 위해 기계학습 방법을 활용하여 단어의 양 및 특정 단어의 빈도수 그리고 단어 간의 연관성 분석을 통해 의료 기사의 적절성 판단을 분류 및 예측하였다. 적절성 판단의 기준은 서울 모 종합병원에서 제공되는, 의료진에 의해 적절성 판단 후 게재된 기사들을 기준으로 하였다. 2014년 전체 데이터를 분류 및 예측하여 5-fold cross validation 방법을 통해 검증한 결과, newscount와 volume이 적절성 예측 모델에 주요 특징으로 나타났고, 분류모델의 정확도는 76~82%를 보이는 것을 알 수 있었다. 또한, 2015년 의료기사를 입력하여 적절성을 예측 및 의료진의 검증 결과, 예측한 기사의 60%가 적절하다고 판단되었고, 제목과 기사의 연관성이 환자에게 적합한 기사를 택하는데 주요한 요인인 것을 알 수 있었다.

본 논문은 기사의 적절성을 판단하는 명확한 기준이 없다는 문제와, 의료기사의 범위가 광범위 하다는 점, 그리고 FALSE로 분류된 데이터가 상당 수를 차지한다는 문제 등의 한계를 지니고 있지만, 하루에도 수없이 보도되는 의료정보에서 환자가 신뢰할 수 있는 정보를 보다 쉽고 빠르게 제공할 수 있다는 점에서 의미있는 연구라고 생각한다. 추후 적절성의 기준을 평가하는 명확한 기준이 정립된다면 정확도를 보다 높일 수 있을 것이라고 생각한다.

또한, 본 논문에서 제시하는 알고리즘을 의료기사가 아닌 보다 다양한 분야에 적용한다면, 수많은 미디어를 통해 제공되는 정보들을 개개인의 니즈에 맞춰 신뢰도 높은 정보만을 분류하여 제공받을 수 있고, 미디어 또한 기사의 신

뢰도를 보다 쉽고 빠르게 검증하여 소비자에게 제공하는데 활용될 수 있을 것이라고 생각한다.

REFERENCE

Kyu-Won Jung, Young-Joo Won, Hyun-Joo Kong, Chang-Mo Oh, Hyunsoon Cho, Duk Hyoung Lee, Kang Hyun Lee(2015), Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2012. Cancer Res Treat, pp. 127-141.

[illegible](2002), [illegible], 10(2).

[illegible](2014), [illegible], 11(4). pp. 169-175.

Braga-Neto,U.M, Zollanvari,A. and Dougherty,E.R(2014), Cross-Validation Under Separate Sampling: Strong Bias and How to Correct it, Bioinformatics, 30, pp. 3349-3355.

[illegible](2013), [illegible]

[illegible](2012), [illegible] PC [illegible], 10(4), pp. 119-128.

[illegible](2011), [illegible]

Rodolfo Passalacqua, Cterina Caminiti, Stefania Salvagni, Sandro Barni, Giordano D. Beretta, Paolo Carlini, Antonio Contu, Francesco Di Costanzo, Lucia Toscano, Francesco Campione(2004), Effect of Media Information on Cancer Patients' Opinions, Feelings, Decision-Making Process and Physician-Patient Communication, American Cancer Society.

Juanne N. Clarke, Michelle M. Everest(2006), Cancer in the mass print media: Fear, uncertainty and the medical model, Social Science & Medicine, 62, pp. 2591-2600.

Jae Woo Nam, Seonghee Kim(2013), A Study on the Effect of Presentation Modes of Health Information on Information Perception, Journal of the Korean Biblia Society For Library And Information Science, 24(4), pp. 217-238.

[illegible](2003), [illegible], [illegible], pp. 68-82.

[illegible](2008), [illegible], [illegible], pp. 158-162

Soojung Kim(2012), An Exploratory Study of Undergraduate Students' Health Information Needs and Seeking Behaviors in Social Media, Journal of the Korean Biblia Society For Library And Information Science, 23(4), pp. 239-260.

[illegible] 2010.12.10. [illegible] [online]. [cited 2012.12.15]. <http://www.yonhapnews.co.kr/economy/2010/12/10/0303000000AKR20101210195000017.HTML?template=2088>.

Kil-Hong Joo, Eun-Young Shin, Joo-Il Lee, Won-Suk Lee(2011), Hierarchical Automatic Classification of News Articles based on Association Rules, Journal of Korea Multimedia Society, 14(6), pp. 730-741.

Jieun Son, Seoung Bum Kim(2014), Rule Selection Method in Decision Tree Models, Journal of the Korean Institute of Industrial Engineers, 40(4), pp. 375-381.

[illegible](2012), [illegible]