KIPS Tr. Software and Data Eng.
Vol.5, No.11 pp.541~548 pISSN: 2287-5905

http://dx.doi.org/10.3745/KTSDE.2016.5.11.541

Linguistic Features Discrimination for Social Issue Risk Classification

Hyo-Jung Oh† · Bo-Hyun Yun†† · Chan-Young Kim†††

ABSTRACT

The use of social media is already essential as a source of information for listening user's various opinions and monitoring. We define social 'risks' that issues effect negative influences for public opinion in social media. This paper aims to discriminate various linguistic features and reveal their effects for building an automatic classification model of social risks. Expecially we adopt a word embedding technique for representation of linguistic clues in risk sentences. As a preliminary experiment to analyze characteristics of individual features, we revise errors in automatic linguistic analysis. At the result, the most important feature is NE (Named Entity) information and the best condition is when combine basic linguistic features. word embedding, and word clusters within core predicates. Experimental results under the real situation in social bigdata - including linguistic analysis errors - show 92.08% and 85.84% in precision respectively for frequent risk categories set and full test set.

Keywords : Risk Detection, Text Classification, Linguistic Feature, Feature Discrimination, Word Embedding

사회적 이슈 리스크 유형 분류를 위한 어휘 자질 선별

오 효 정† · 윤 보 현†† · 김 찬 영†††

요 약

사용자의 다양한 의견을 수렴하고 모니터링하기 위한 정보원으로써 소셜미디어의 활용은 이미 필수가 되었다. 본 논문은 소셜미디어에 나타난 다양한 이슈 중 여론 형성에 악영향을 끼치는 부정적 사건을 이슈 '리스크'로 정의, 그 세부 유형을 자동으로 분류하는 모델을 개발하고자 한다. 이를 위해 소셜미디어에 나타난 다양한 어휘 자질을 선별, 그 효과를 규명하였다. 특히 리스크 문장의 어휘 구문 특징을 표현하기 위한 자질로 워드 임베딩 학습 결과를 활용한다. 개별 어휘 자질의 특징을 분석하기 위해 언어분석 오류를 보정한 환경에서 수행한 실험 결과, 가장 효과가 큰 자질은 개체명 자질로 분석되었으며, 기본 어휘 자질을 기반으로 주요 술부의 워드 임베딩 결과와 워드 클러스터 결과를 모두 조합한 경우가 최고 성능을 보이는 것으로 파악되었다. 실제 소셜빅데이터에 적용하는 환경과 유사하도록 자동 언어분석 결과의 오류를 포함한 조건에서 실험한 결과, 고빈도 평가셋에서는 92.08%의 성능을, 전체 58개 범주 평가셋에서는 85.84%의 성능을 얻었다.

키워드 : 리스크 탐지, 문서 분류, 어휘 자질, 자질 선별, 워드 임베딩

1. 서 론

사용자의 다양한 의견을 수렴하고 모니터링하기 위한 정보원으로써 소셜미디어의 활용은 이미 필수가 되었다. 특히 소셜빅데이터를 통해 공공 여론의 동향을 파악하고 이를 선제적 대응 매체로 활용하는 연구들이 매우 활발히 진행 중이다. 최근에는 이와 관련된 프로젝트가 환경탐색(Horizon Scanning)이라는 이름으로 국가적인 차원에서 진행되고 있는데, 싱가포르의 RAHS(Risk Assessment Horizontal Scanning) 프로젝트, 영국의 호라이즌 스캐닝 센터(The Foresight Horizon Scanning Center), EU의 FutureICT, iKnow 프로젝트가 대표적이다[1]. 이들 프로젝트는 환경탐색을 통해 획득한 데이터를 분석하여 국가적 영향을 미칠 수 있는 잠재적 위험요소와 불확실성 요소를 탐색하고 이머징 이슈를 분석함으로써, 각국의 미래 시나리오를 연구하고 대응방안을 마련하기 위한 기초자료로 활용하는데 그 목적이 있다.

국내에서도 이와 같이 여론 동향을 파악하고 의사결정을 지원하기 위해 소셜미디어를 활용한 연구가 다수 발표되었다. 초반에는 주로 기업과 기관을 중심으로 상품 브랜드 인식 수렴에 대한 연구가 주를 이루었으나 2012년 총선을 기점으로 소셜미디어 상의 민심 동향을 보여 줄 수 있는 선거 관련 서비스 사례 등이 두드러지고 있다[2]. 그러나 이들 서비스는 단기

※ 이 논문은 2016년도 전북대학교 연구기반조성비지원에 의하여 연구되었음.
※ 이 논문은 2016년도 한국연구재단 연구비 지원에 의한 결과의 일부임 (과제번호: 2016R1A2B1008000)
† 정 회 원 : 전북대학교 대학원 기록관리학과 조교수
†† 정 회 원 : 목원대학교 컴퓨터교육과 교수
††† 비 회 원 : 전북대학교 의학전문대학원 부교수
Manuscript Received : October 4, 2016
Accepted : October 12, 2016
* Corresponding Author : Chan-Young Kim(happyhill@jbnu.ac.kr)

간 내에 화제가 된 사건을 중심으로 단편적인 정보 분석결과를 제공하는 것으로, 대량의 소셜미디어를 분석하여 이를 시계열에 따라 모니터링해주는 고차원 리스닝 플랫폼 서비스로의 확장은 아직 초기 단계에 머무르고 있다[3].

본 연구팀에서는 최종적으로 소셜빅데이터의 내용을 분석함으로써 잠재적 위험이 될 수 있는 이슈를 자동으로 모니터링하는 방법을 고안하고자 한다. 그 첫걸음으로 사회적으로 파장이 큰 이슈 중 공공여론이 부정적으로 형성될 이슈를 '리스크(risk)'로 정의, 세부 유형을 자동으로 분류하는 모델을 제안한다. 특히 본 논문에서는 소셜미디어에 나타난 어휘 자질의 특징을 분석하고, 이를 기계학습 방법에 적용함에 있어 리스크 탐지를 위한 학습 대상 자질의 유형을 정의, 그 효과를 분석하고자 한다.

2. 관련 연구

본 논문에서 다루고자 하는 문제는 소셜빅데이터에 다수 사용자로부터 제기된 다양한 이슈 문장들 중 사회적으로 파장이 크면서 공공여론이 부정적으로 형성될 문장을 리스크로 판별, 그 유형을 분류하는데 있다. 이러한 문제는 기존의 문서분류(Text Classification, 이하 TC)와 매우 유사하다.

TC 문제를 해결하기 위해 기계학습 방법을 적용한 연구들은 다수 진행되어 괄목할 만한 성능을 보이고 있다. 가장 두각을 나타내고 있는 지도 학습(supervised learning) 방법으로는 SVMs(Supported Vector Machines)[4, 5]이나 CRFs (Conditional Random Fields)[6]를 이용한 연구들이 발표되었으며, 최근에는 심층 학습(Deep Learning, 이하 DL)[7, 8]을 이용한 방법들도 시도되고 있다. 이번 장에서는 이러한 기계학습을 적용하기 위한 자질(feature) 선정과 관련된 연구 동향을 살펴보도록 한다.

2.1 최적 자질 선정

TC에 기계학습을 적용하기 위해서는 대상문서집합의 특성을 대표할 수 있는 자질을 규명하고 이를 선별하여 양질의 자질만을 학습 대상으로 선정하는 작업이 필요하다. 기존의 자질 선별과 관련된 연구 동향을 살펴보면, [9]는 특정 단어가 가지는 가중치는 시기에 따라 달라진다는 점을 이용하여 날짜 정보를 추가하였고, [10]은 단어로 구성된 기본 특징에 단어 간의 관계정보를 추가하고 이를 확장하여 얻어낸 혼합 특징을 사용하여 성능향상을 꾀하였다.

SVM을 TC에 적용한 연구들도 다양한 분야에서 시도되었다. [4]에서는 다양한 자질을 사용하여 SVM 분류 성능의 영향을 확인하였는데, 특히 트윗(tweet) 문서의 감정 분류를 위해서는 이모티콘과 사용자의 극성 자질이 가장 큰 영향을 미치는 것으로 분석하였다. [5]는 언어 종속적 단위인 형태소 자질어와 언어 독립적 단위인 n-gram 자질어 그리고 이들을 조합한 복합 자질어 집합을 대상으로 각각 한국어와 중국어로 작성된 인터넷 신문기사를 SVM으로 분류하는 실험을 수행하였다.

그 밖에도 [11]은 자유롭게 기술되는 비격식 문서를 SVM을 활용해 분류함에 있어 LDA(Latent Dirichlet allocation) 단어 분포를 사용하여 자질(feature)을 교정하고 확장하는 방법을 제안하였다. 본 논문에서는 소셜미디어의 내용을 분석하기 위해 적용된 언어분석 수행 단계별로 어휘 자질을 선정, 그 효과를 분석하여 최적 조합을 선정한다.

2.2 워드 임베딩

TC에서 문장 내에 출현한 단어의 의미를 파악하기 위한 방법은 매우 중요하다. 워드 임베딩(word embedding)이란 단어가 가지는 의미나 맥락을 고려하여 단어를 벡터로 표현한 것으로, 학습 입력 문장 내에 있는 단어와 인접 단어의 관계를 이용해 단어의 의미를 학습한다. 주로 신경망 언어모델(Neural Network Language Model: NNLM) 기반 접근 방법이 최근 연구 경향으로 볼 수 있는데, 2003년 Bengio[7]는 NNLM에서 각각의 어휘를 분산된 연속 값으로 인코딩한 벡터를 활용한 언어모델 구축 방법을 제안하였다. 이후 구글의 Mikolov[8]는 재귀적 신경망 언어모델(Recursive NNLM)을 활용해 계산 속도를 획기적으로 줄인 방법을 제안, word2vec이라는 공개툴[12]로 제공되고 있다. word2vec의 학습 방법에는 2가지 종류가 있는데, CBOW(Continuos Bag Of Words) 방식은 주변 단어가 만드는 맥락을 이용해 타겟 단어를 예측하는 것이고 skip-gram은 한 단어를 기준으로 주변에 올 수 있는 단어를 예측하는 방식이다. 일반적으로 대규모 데이터 셋에서는 skip-gram이 더 정확한 것으로 알려져 있으며 본 연구에서도 이 방식을 차용하였다.

워드 임베딩 기술을 한국어 분석에 활용한 경우도 다수 있다. [6]은 개체명 인식을 위한 자질로 워드 임베딩에 기반한 어휘 클러스터를 활용하여 성능향상을 꾀하였고, [13]에서는 지식 추출을 위한 트리플 생성시 의미 유사도 계산에 워드 임베딩 기술을 활용하였다. 본 논문에서는 리스크 문장 내의 부정적 의미를 암시하는 구문을 학습하기 위한 자질로, 예를 들면 '된서리를 맞았다'나 '먹구름이 드리울 전망이다'와 같은 다양한 형태의 부정 숙어 표현을 그룹핑하기 위해 워드 임베딩 결과를 활용한다.

3. 사회적 이슈 리스크 유형

본 논문에서 탐지하고자 하는 '사회적 이슈 리스크'란 소셜 빅데이터로부터 사회적으로 파장이 큰 이슈 중 공공여론이 부정적으로 형성될 이슈를 말한다. 여기서 '리스크'란 소셜웹 미디어에서 발생한 다양한 이슈 중에서 이슈 대상 주체(target)에게 위협이 될만한 혹은 잠재적으로 위험을 내포

하고 있는 사건을 의미한다. 예를 들면 공공의 공통관심 이슈로 인물이나 기관의 부패 혹은 보건, 복지 관련 정책이나 국민안전사고 등으로, “최태원 검찰수사 개시, SK 주가폭락” 뉴스나 “선관위 홈페이지 또 디도스 공격”과 같은 사건을 들 수 있다.

본 연구팀에서는 사전연구[14]를 통해 소셜미디어에 나타난 사회적 이슈 리스크의 유형을 분석하였다. 분석 대상 소셜미디어로는 뉴스미디어와 트위터를 선정, 19대 총선과 18대 대선 등 정치·사회적 이슈가 많이 발생한 2012년도를 전수조사대상 기간으로 선정하였다. 각각 1,437,188건의 뉴스와 1,002,240,097 트윗을 수집하였으며 조사 결과, 평균적으로 한 달에 생성되는 기사 중 정치·사회적으로 부정적인 영향을 미친 사건, 사고를 다룬 기사가 전체 37.9%에 해당하는 것을 알 수 있다. 이 중에서 처음 사건이 보도된 이후 해당 사건에 관해 중복으로 다룬 기사를 제외한 리스크 후보 대상 문서 수는 월별 평균 1,920건으로, 전체 기사 중 1.6%에 해당하는 것으로 분석되었다. 이는 하나의 사건당 평균 23.65개의 후속보도가 다루어졌음을 의미한다.

Table 1은 사전연구[14]를 통해 최종 정의된 사회적 이슈 리스크의 세부 유형을 정리한 것으로, 최종 선정된 리스크 유형은 전체 수집된 리스크 문서 중 리스크 위험 대상이 명확한 사건을 대상으로 선정하였다. 본 연구의 궁극적인 목적은 소셜미디어로부터 수집된 문서에 사회적 이슈 리스크의 세부 유형을 부여하는 것으로, 예를 들면 “기상청, 납품단가 조작 의혹”이라는 기사는 [공공기관_위법행위] 유형으로, “공정위, 애플·구글 ′특허횡포′ 손본다”라는 기사는 [IT업체_조사]라는 유형으로 분류된다.

Table 1. Social Risk Category[14]

Domain	# of Classes	Classes
Public Organization	11	갈등, 경쟁, 법적조치, 법정판결, 부정여론, 소송, 시위, 위법행위, 정보유출, 제재, 조사
Person	12	갈등, 건강악화, 경쟁, 법적조치, 법정판결, 부정여론, 사퇴요구, 소송, 시위, 위법행위, 정보유출, 조사 , 징계
Food	9	가격상승/하락, 리콜, 부정여론, 불매운동, 위법행위, 유해식품, 제재, 판매감소
Automobiles	6	가격하락, 결함, 경쟁, 리콜, 부정여론, 판매감소
IT Company	12	갈등, 경영위기, 경쟁, 범정판결, 부정여론, 불매운동, 사업종료, 소송, 위법행위, 정보유출, 조사, 제재
Smart Devices	8	가격하락, 결함, 경쟁, 부정여론, 정보유출, 제재, 판매감소
6	58	

4. 사회적 이슈 리스크 어휘 자질 후보

어휘 자질(linguistic feature)은 미디어 내용 분석에 있어 가장 기본이 되는 자질로, 본 논문에서는 한국전자통신연구원(이하 ETRI)에서 개발한 어휘의미분석기술[15]을 활용해 분석 단계별로 어휘 자질을 추출하였다.

4.1 단어 자질

문장별 리스크 탐지를 위한 가장 기본적인 자질로 형태소와 어휘의미 정보를 선정하였다. Fig. 1은 “최태원 SK 회장, 검찰수사로 7천410억원 허공으로”라는 기사에서 발췌한 문장에 대해 ETRI 형태소 분석기와 어휘의미태깅(WSD: Word Sense Disambiguate) 모듈을 적용한 결과로, WSD 결과의 어휘의미코드는 표준국어대사전의 의미표지를 따른다. 본 논문에서는 형태소 태그와 의미코드 모두를 단어 자질로 활용하였다.

최태원 SK 회장이 검찰수사가 시작된 후 무려 7천410억원을 허공에 날린 것으로 전해졌다

어절	형태소 분석	WSD 결과
최태원	최태원/NNP	00
SK	SK/SL	00
회장이	회장/NNG 이/JKS	07 00
검찰수사가	검찰/NNG 수사/NNG 가/JKS	02 18 00
시작된	시작되/VV ㄴ/ETM	01 00
후	후/NNG	08
무려	무려/MAG	02
7천410억원을	7/SN 천/NR 410/SN 억/NR 원/NNB 을/JKO	00 03 00 04 01 00
허공에	허공/NNG 에/JKB	00 00
날린	날리/VV ㄴ/ETM	02 00
것으로	것/NNB 으로/JKB	01 00
전해졌다.	전하/VV 어/EC 지/VX 었/EP 다/EF ./SF	00 00 04 00 00 00

Fig. 1. An example of Morphology Analysis and WSD

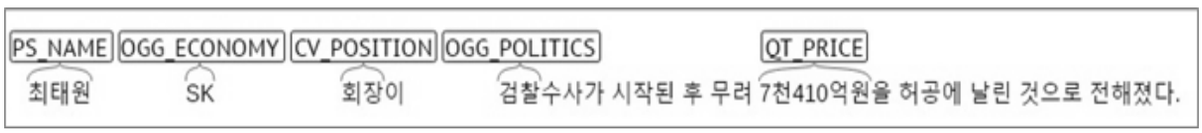


Fig. 2. An Example of Named Entity Recognition

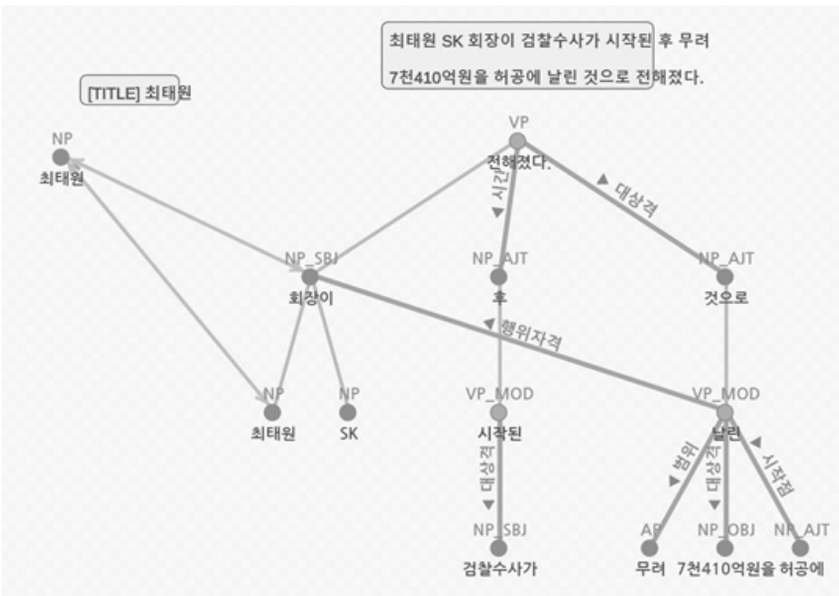


Fig. 3. An example of Parsing and Semantic Role Labeling

4.2 개체명 자질

3장에서 기술한 사전조사[14]을 통해 사회적 이슈 리스크를 보도하고 있는 문장의 경우, 해당 리스크의 대상(target)이 되는 개체(entity)가 언급되는 경우가 대부분인 것을 알 수 있었다. 일부 개체가 나타나지 않은 문장도 발견되었으나, 이는 주로 뉴스 보도문이나 위키피디아 설명문에서 자주 나타나는 것으로 파악되었다. 이러한 문장들의 특성은 바로 앞 문장에서 언급한 주어나 목적어를 생략하는 경우가 대부분이었으며 이는 주어복원(zero anaphora resolution)[15]을 통해 복구가 가능하였다. 이와 같은 특성을 반영하여 개체명 분석 결과를 두 번째 어휘 자질로 사용하였다.

Fig. 2는 Fig. 1의 예시 문장에 대한 개체명 인식 결과를 도식화한 것으로, '최태원'은 인명(PS_NAME)으로, 'SK'와 '검찰'은 각각 기업(OGG_ECONOMY)와 정부기관(OGG_POLITICS)라는 조직명, 마지막으로 '7천410억원'은 가격(QT_PRICE)으로 인식되었다. 본 논문에서는 실제 어휘정보는 단어 자질로, 개체명 태그 정보는 개체명 자질로 나누어 학습하였다. 학습에 사용한 개체명 태그는 최상위 15개, 세부 146개이다[15].

4.3 문장 구조 자질

앞서 설명한 바와 같이, 이슈 리스크를 보도하고 있는 문장의 경우에는 리스크 대상에 미치는 영향이나 리스크 대상이 취한 행위에 관한 보도가 대부분이다. 따라서 구문분석을 통해 문장 구조를 분석하고 의미역(semantic role)을 파악함으로써 리스크 구문 구조의 특성을 학습한다. 또한 술어-논항(predicate-argument) 관계를 통해 4.4절에서 설명할 관용적 표현을 인식하기 위한 학습 범위로 설정한다. Fig. 3은 Fig. 1의 예시 문장의 구조 분석 결과로, 주어인 '최태원' 회장이 7천410억원을 '허공에 날린' 대상임을 나타내며, 이는 [인물_조사] 리스크 분류의 단서로 활용 가능하다.

4.4 워드 임베딩을 통한 감성 자질

리스크 탐지에 사용한 네 번째 자질은 문장에 드러난 감성(sentiment) 정보이다. 3장에서 기술한 사회적 이슈 리스크의 정의를 짚어보면, 사회적으로 파장이 크며 앞으로 여론이 부정적으로 형성될 사건이나 사고를 의미한다고 밝혔다. 여기서 '부정적'인 사건/사고를 판단하는 기준은 뉴스 문장의 구문적 특성과 논조에 따라 판별될 수 있다. 예를 들면 '물의를 일으키다 / 구설수에 오르다 / 문제로 분석 되었습니다' 등과 같이 직접적으로 문제가 있음을 밝힌 경우 혹은 '파장이 일고 있다 / 시민들이 불안에 떨고 있다 / 심각히 우려되는 상황입니다' 등의 어구가 나타난 문장이 사회적 이슈 리스크를 보도하고 있는 문장으로 파악되었다.

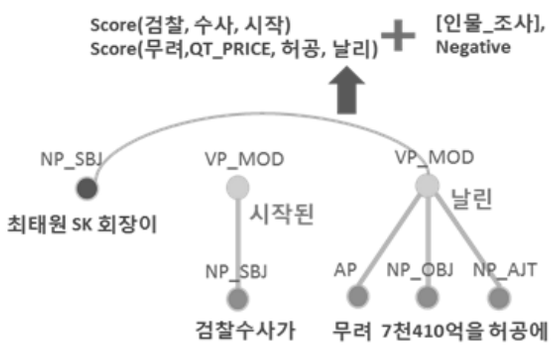


Fig. 4. An Example of Word Embedding

앞 절에서 기술한 세 가지 자질 유형은 전통적인 언어분석과정에서 추출된 결과로, 단일(복합명사 포함) 어휘에 대한 분석 계층에 따른 유형이다. 이에 비해 감성 자질은 숙어나 관용적 표현에 대한 이해가 필수적이다. 기계 학습을 기법을 통해 이와 같은 지식을 자동으로 습득하기 위해서는 매우 정교하고 많은 양의 학습데이터가 필요하다. 이와 같은 단점을 보완하기 위해 본 논문에서는 뉴스보도문에 자주 나타나는 부정적 표현 어구 벡터(word vector)를 skip-gram 기반 워드 임베딩 기법[8]을 통해 계산하였다. 이후 생성된 1-gram 에 대해 K-means 클러스터링을 수행, 생성된 클래스 번호를 리스크 감성 어휘 자질로 활용하였다.

Skip-gram 학습을 위한 입력은 Fig. 1의 형태소 단위이되, 가격(QT_PRICE)이나 거리(QT_LENGTH) 등 수치(QT 계열)를 나타내는 개체명의 경우, 실제 어휘 대신 개체명 태그로 대치하였다. 전체 문장내의 모든 단어열을 입력대상으로 기초 실험을 수행한 결과, 많은 차원의 어휘벡터가 생성되어 임베딩 클러스터의 수 역시 많이 생성되었다. 이는 결과적으로 리스크 분류를 위한 자질 클러스터 수가 많아지는 것으로, 학습에 부정적인 영향을 끼친다.

이러한 점을 보완하여 학습 입력 단어열을 문장 내 모든 어휘열이 아닌 핵심 술부(core predicates) 구문 트리 내로 한정하여 임베딩 자질로 사용하였다. Fig. 4는 [인물_조사] 리스크에 해당하는 문장을 예시한 것으로, 최태원 회장에 대한 검찰 조사로 인해 SK 주가가 떨어진 사건을 보도한 문장이다. 여기서 최상위 술부인 '전해졌다'의 노드는 뉴스보도문의 관용적 특성으로, 특별한 의미가 없는 술어라 제외하고 '시작된'과 '날린'에 연결된 구문의 형태소 열만을 입력치로 사용, [인물_조사]에 해당하는 부정적 의미를 표현하는 자질로 학습된다.

5. 사회적 이슈 리스크 분류를 위한 자질 조합

5.1 실험집합

학습 및 평가에 사용한 실험집합은 사전조사를 통해 수집된 리스크 후보 문서 23,034개 중 Table 1에 정의된 대상 리스크 유형에 해당하는 문서 3,000개를 임의로 선정, 문장별 이슈 리스크 분류를 태깅하였다. 그 결과 총 24,385 문장 코퍼스가 구축되었으며, 이를 2:1의 비율로 나누어 학습 및 평가에 사용하였다. 특히 개별 어휘 자질 효과 분석을 위해 학습문장 중 일부를 선별, 언어분석 결과의 오류를 보정하여 평가하였다.

실험에 사용한 기계학습 기법은 SVM 알고리즘[14]을 적용하였으며, 워드 임베딩 학습은 공개 소스인 word2vec[12]를 이용하여 학습 후 클러스터링을 수행, 생성된 클래스 번호를 자질로 활용하였다. Table 2는 실험에 사용한 코퍼스의 통계를 보여준다.

Table 2. Experiment Corpus Statistics

Attributes	# of values
Documents	3,000
Risk Target Sentence	24,385
Training Set (Controlled Set)	16.210 (1,621)
Test Set (Top-20 Risk Set)	8,175 (3,713)
Input Words	443,807
Input Tokens	1,239,679

5.2 기초 실험

본 논문의 주요 학습대상은 어휘 자질로, 언어분석 성능에 따라 최종 리스크 탐지 성능도 영향을 받는다. 제안된 어휘 자질 유형이 리스크 분류에 미치는 영향을 파악하기 위해 먼저 언어분석 결과의 오류가 없는 상황에서의 실험을 수행하였다. 이를 위해 학습 데이터 중 10%를 선별, 1,621문장의 시스템 언어분석 결과에 대한 수작업으로 검증, 보정하였다.

Table 3은 자질 사용에 따른 분류 성능을 보여준다. 가장 기본 자질인 형태소 태그와 의미코드를 사용한 경우(Word) 87.72%의 정확도를 보인 반면, 개체명 자질을 포함해 학습한 결과(Word+NE)는 92.60%의 정확도를 보여 5.56%의 성능 향상을 얻을 수 있었다. 문장구조 자질(Struct)을 포함한 경우, 개체명까지만 사용한 경우에 비해 정확도가 93.71%로 향상되었으나 그 효과가 1.2%로 매우 비미하였다. 이는 리스크 문장에서의 주어나 목적어가 개체명인 경우가 대부분으로, 이미 문장구조 자질 활용 효과를 나타낸 것으로 해석될 수 있으며 나아가 리스크 탐지에 있어 개체명 자질이 매우 중요함을 의미한다.

워드 임베딩을 위해 문장 전체 단어열을 입력으로 한 경우(WE full)와 핵심 술부내의 단어열만을 입력으로 한 경우(WE core)의 성능을 비교한 결과, 핵심 술부 구조로 제한한 경우(94.51%)가 전체를 대상으로 학습한 경우(93.89%)에 비

해 높았다. 특히 문장 전체 학습 결과는 문장구조 자질까지만 활용한 경우에 비해 성능 향상이 거의 없었다(0.13%). 또한 워드 임베딩 벡터자질로부터 보다 의미 있는 어휘 집합(chunk)를 추출하기 위해 K-means 알고리즘을 활용해 클러스터링을 수행, 상위 300개 클래스 번호를 자질로 활용하였다. 위 분석 결과를 종합해보면, 가장 효율적인 자질 조합으로는 기본 어휘 자질(Basic)에 주요 술부 내의 워드 임베딩 결과(WE core)와 워드 클러스터 결과(WC 300)을 활용한 경우(Model 6)로, 최종 94.69%의 정확도를 보였으며, 이는 오류가 없는 완벽한 언어분석 모듈을 활용한 경우의 최대 기대치를 의미한다.

Table 3. Preliminary Experiment Results

Model	Methods	Precision	Improvements
1	Word(Morp. + WSD)	87.72%	
2	Word+ NE	92.60%	5.56% over Model 1
3	Basic(Word+ NE+Struct.)	93.71%	1.20% over Model 2 6.82% over Model 1
4	Basic + WE full	93.83%	0.13% over Model 3 6.82% over Model 1
5	Basic + WE core	94.51%	0.73% over Model 4 7.74% over Model 1
6	**Basic + WEcore + WC 300**	**94.69%**	0.20% over Model 5 1.05% over Model 3 **7.95% over Model 1**

5.3 사회적 이슈 리스크 분류

앞서 설명한 기초실험은 언어분석 성능이 100%일 때를 가정하고 어휘 자질 자체의 효과분석을 수행하였다. 이번 실험에서는 실제 소셜빅데이터로부터 리스크를 탐지하는 환경에서의 실험을 수행하고자 한다.

Table 4는 본 논문에서 제안한 자질 조합별 리스크 분류 성능을 정리한 결과이다. 이는 언어분석의 오류를 포함하는 성능으로, 실제 소셜 빅데이터를 대상으로 적용했을 때의 체감 성능으로 간주할 수 있다. 전체 58개 리스크 분류에 해당하는 평가셋은 8,175개이나 이 중 고빈도 리스크 분류 상위 20위에 해당하는 문장은 3,713개로, 고빈도 집합에 대한 성능 분석을 먼저 수행하였다.

고빈도 집합에 대한 실험 결과, 어휘자질(단어+개체명+문장구조, Basic)만 활용한 경우 88.61%의 정확도를 보인 반면 워드 임베딩 자질을 활용한 경우 다소 높은 성능(90.01%)을 얻을 수 있었다. 한편, 워드 클러스터만을 활용한 경우, 어휘자질만을 활용한 경우에 비해 미미하게 증가했지만(0.7%) 워드 임베딩 벡터값을 그대로 활용한 경우에 비해서는 다소 감소하였다(-0.87%). 그러나 최종적으로 모든 자질을 종합해 사용한 경우는 워드 임베딩만 사용한 경우와 워드 클러스터를 단독으로 사용한 경우에 비해 3.92%가 증가한 92.08%의 성능을 보였다. 이러한 결과는 5.2절의 최대 기대치 성능(94.69%)과 비교해 언어분석기 오류를 포함한 결과임에도 불구하고 2.61%의 차이밖에 나타나지 않는 것으로, 다양한 자질의 조합이 오류를 상쇄하는 것으로 유추할 수 있다.

Table 4. Social Risk Classification Results

Model	Methods	Precision	Improvements
7	Basic(Word+ NE+Struct.)	88.61%	
8	Basic + WE core	90.01%	1.58% over Model 7
9	Basic + WC 300	89.23%	-0.87% over Model 8 0.70% over Model 7
10	**Basic + WE core + WC 300** (Top-20 Risk set)	92.08%	3.20% over Model 9 3.92% over Model 7 -2.61% over Model 6
11	Basic + WE core + WC 300 (full Test)	85.84%	-6.24% over Model 10

전체 58개 리스크 분류셋을 대상으로 실험한 결과(Model 5), 85.84%의 정확도를 보였다. 이와 같은 결과는 고빈도 집합 성능(Model 4)과 매우 큰 차를 보이는데, 이는 문서집합이 고빈도 리스크 분류에 해당하는 예제가 다수 포함되어 학습에 영향을 미쳤기 때문으로 판단된다. 또한 시스템 언어분석 오류 역시 고빈도 어휘나 개체명을 포함한 문장에서 발생빈도가 적었기 때문으로 해석된다. 이와 같은 현상은 Table 3과 Table 4의 개선도를 비교해보면 알 수 있는데, 언어분석 오류가 통제된 환경에서 기본 어휘자질만 사용한 경우에 비해 워드 임베딩 벡터와 클러스터 결과를 모두 조합한 경우의 개선 효과는 1.05%(Model 3 → Model 6)인 반면, 언어분석 오류를 포함한 환경에서는 3.92%(Model 7 → Model 10)가 증가해 개선 효과가 더 두드러짐을 알 수 있었다.

6. 결 론

본 논문은 소셜미디어에 나타난 다양한 이슈 중 여론 형성의 악영향을 끼치는 부정적 사건을 이슈 리스크로 정의, 그 세부 유형을 자동으로 분류하기 위해 다양한 어휘 자질을 선별하고 그 효과를 규명하였다.

언어분석 오류를 보정한 환경에서 실험한 결과, 가장 효과가 큰 자질은 개체명 자질로 분석되었다. 이는 리스크 문장의 특성상 리스크의 대상 혹은 리스크를 발생시킨 주체가 대부분 인명이나 기관, 제품 등 특정 개체명인 경우가 대부분이기 때문으로 해석될 수 있다. 이와 더불어 워드 임베딩을 통해 의미있는 어휘 클러스터를 사용하는 경우가 가장 높은 정확도를 보였다.

분석된 자질 조합을 기준으로 자동 언어분석 결과의 오류를 포함한 환경에서 수행한 실험 결과, 고빈도 리스크 셋에서는 92.08%의 성능을, 전체 58개 분류셋에서는 85.84%의 성능을 얻었다. 언어분석 결과의 보정된 결과에 비해 전체적으로 정확도는 낮은 성능을 보였으나, 워드 임베딩 벡터와 클러스터 결과를 모두 조합한 경우의 효과는 더 큰 것으로 분석되었다. 이는 언어분석 오류가 다수 발생하는 소셜 빅데이터 환경에서 보다 강건한 모델임을 유추할 수 있다.

이슈 리스크의 특성상 부정적 표현 어구에 대한 인식이 매우 중요하다. 향후 연구방향으로는 현재의 워드 임베딩 결과로부터 숙어나 관용표현 사전을 반자동으로 구축하는 연구를 진행하고자 한다. 나아가 본 논문에서 정의한 소셜 미디어에 나타난 어휘 자질 이외에도 소셜미디어가 갖는 다양한 메타데이터 및 특성을 활용한 자질을 발굴, 이를 활용하고자 한다.

References

[1] G. H. Kim, S. Trimi, and J. H. Chung, "Big-data applications in the government sector," *Communications of the ACM,* Vol.57, No.3, pp.78-85, 2014.

[2] C. H. Lee, J. Hur, and H. J. Oh, et al., "Technology Trends of Issue Detection and Predictive Analysis on Social Big Data," *Electronics and Telecommunications Trends*, Vol.28, No.1, pp.62-71, 2013.

[3] J. Hur, C. H. Lee, and H. J. Oh, et al, "Automatic Generation of Issue Analysis Report Based on Social Big Data Mining," *Korea Information Science Society (KISS) Journals*, Vol.3, No.12, pp.553-564, 2014.

[4] C. H. Hong and H. S. Kim, "Comparative Study of Various Machine-learning Features for Tweets Sentiment Classification," *Korea Contents*, Vol.12, No.12, pp.471-478. 2012.

[5] M. Y. Ren and S. J. Kang, "Comparison Between Optimal Features of Korean and Chinese for Text Classification," *Journal of Korean Institute of Intelligent Systems*, Vol.25, No.4, pp.386-391, 2015.

[6] Y. S. Chio and J. W. Cha, "Korean Named Entity Recognition and Classification using Word Embedding Features," *Journal of KIISE*, Vol.43, No.6, pp.678-685, 2016.

[7] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Journal of Machine Learning Research*, Vol.3, pp.1137-1155, 2003.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Proceedings of the ICLR Workshop,* 2013.

[9] B. Shim, J. Park, and J. Seo, "Term Weighting Using Date Information and Its Appliance in Automatic Text Classification," in *Proceedings of the 19th Annual Conference on Human and Cognitive Language Technology*, Vol.10, pp.169-173, 2007.

[10] J. In, J. Kim, and S. Chae, "Combined Feature Set and Hybrid Feature Selection Method for Effective Document Classification," *Journal of Korean Society for Internet Information*, Vol.14, No.5, pp.49-57, 2013.

[11] H. K Lee, S. Yang, and Y.J. Ko, "Feature Expansion based on LDA Word Distribution for Performance Improvement of Informal Document Classification," *Journal of KIISE*, Vol. 43, No.9, pp.1008-1014, 2016.

[12] Word2vec [Internet], https://code.google.com/p/word2vec/.

[13] H. G Yoon, S. J. Chio, and S. B. Park, "Improving The Performance of Triple Generation Based on Distant Supervision By Using Semantic Similarity," *Journal of KIISE*, Vol.43, No.6, pp.653-661, 2016.

[14] H. J. Oh, S. J An, and Y. Kim, "Social Issue Risk Type Classification based on Social Bigdata," *Jounrnal of the Korea Contents Association*, Vol.16, No.8, pp.1-9, 2016.

[15] S. J. Lim, C. K. Lee, and D. Y. Ra, "Dependency-based semantic role labeling using sequence labeling with a structural SVM," *Pattern Recognition Letters*, Vol.34, No.6, pp.696-702, 2013.

오 효 정

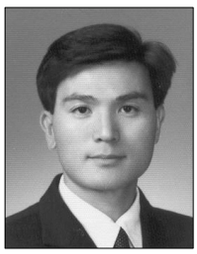
e-mail : ohj@jbnu.ac.kr

2008년 한국과학기술원 컴퓨터공학과 (공학박사)

2000년~2015년 한국전자통신연구원 지식마이닝연구실 책임연구원

2015년~현 재 전북대학교 대학원 기록관리학과 조교수

관심분야 : 정보검색, 질의응답, 빅데이터정보처리



윤 보 현

e-mail : ybh@mokwon.ac.kr

1999년 고려대학교 컴퓨터학과(이학박사)

1999년~2002년 한국전자통신연구원 선임연구원(팀장)

2003년~현 재 목원대학교 컴퓨터교육과 교수

관심분야 : 자연어처리, 정보검색, 시맨틱웹, e-Learning

김 찬 영

e-mail : happyhill@jbnu.ac.kr
1995년 전북대학교 의과대학(학사)
2000년 전북대학교 의과대학원(석사)
2007년 전남대학교 의과대학원(박사)
2005년~현 재 전북대학교
의학전문대학원 부교수
관심분야 : Medical Informatics, Bio Bigdata Processing, Social Risk