Journal of the KIMST, Vol. 21, No. 1, pp. 86-93, 2018
ISSN 1598-9127
DOI http://dx.doi.org/10.9766/KIMST.2018.21.1.086

Research Paper 정보 · 통신 부문

국방 기사 자동 분석 시스템 구축 방안 연구

김현중[1] · 김우주*,[1]

[1] 연세대학교 산업공학과

A Study on Automatic Analysis System of National Defense Articles

Hyunjung Kim[1] · Wooju Kim*,[1]

[1] *Department of Industrial Engineering, Yonsei University, Korea*

(Received 25 September 2017 / Revised 20 December 2017 / Accepted 26 January 2018)

ABSTRACT

Since media articles, which have a great influence on public opinion, are transmitted to the public through various media, it is very difficult to analyze them manually. There are many discussions on methods that can collect, process, and analyze documents in the academia, but this is mostly done in the areas related to politics and stocks, and national-defense articles are poorly researched. In this study, we will explain how to build an automatic analysis system of national defense articles that can collect information on defense articles automatically, and can process information quickly by using topic modeling with LDA, emotional analysis, and extraction-based text summarization.

Key Words : Latent Dirichlet Allocation(잠재 디리클레 할당), Text Mining(텍스트 마이닝), Text Summarization(문서 요약), Sentimental Analysis(감성 분석)

1. 서 론

뉴스미디어는 국민들에게 논란의 여지가 있는 특정 요소를 강조함으로써 이슈화를 통해 대중 여론을 형성할 수 있으며[1], 인터넷과 SNS의 발달은 국민들에게 손쉽게 뉴스를 접할 수 있는 촉매제가 되었다.

언론에서 보도하는 기사는 이슈화되어 대중들에게 여러 방향으로 영향을 미친다는 것은 기존 연구를 통해 밝혀진 바 있으며[2], 따라서 군 입장에서는 군 관련 기사를 빠르고 정확하게 확인하는 것이 대국민 신뢰도를 확립하는데 중요한 과업이 되었다.

군과 관련된 기사는 국방부 대변인실과 각 군에서 제공하는 공보 내용 이외에도 다양한 경로를 통해서 언론사가 취재·보도하기 때문에 어떤 내용을 무슨 논조로 보도할지 확인하는 것은 쉽지 않은 일이다.

또한, 한국언론진흥재단에서 제공하는 2015년 일간 신문 발행부수 현황에 포함된 신문사는 161개사, 2017. 8.17. 기준 네이버에서 제공하는 언론사는 일간지 포함 671개사이며, 2015. 1. 1. ~ 2017. 8.17. 네이버 뉴스 페

* Corresponding author, E-mail: wkim@yonsei.ac.kr

이지 국방/외교 범주에 포함된 기사는 일일 평균 152건 보도로 그 수가 많다.

이렇듯 수많은 매체 속에서 보도 되는 국방 관련 기사를 사람이 별다른 시스템 없이 수동으로 확인하고 분석하는 일은 어려운 일이다.

한편, 학계에서는 정치적 사안에 대한 언론사의 프레임 접근법과 주가 예측모델 개발을 위한 증권 관련 기사 분석 등 정치, 경제 등의 기사 분야에 대한 많은 연구가 이루어져 온 반면, 국방과 관련된 기사를 분석하는 연구는 저조한 실정이다.

Gam et. al.은 경향신문, 한겨레, 동아일보 기사를 사회, 정치 등 6개 분야로 구분하여 분류와 군집 등 텍스트마이닝을 이용 분석한 결과 각 이슈에 대해 신문사별로 다른 프레임으로 긍·부정을 표현하며 자주 사용하는 단어에도 차이가 나는 것을 확인하였으나 국방 분야에는 적용되지 않았다[3].

Cheon et. al.은 국내 주가지수 상승·하락폭이 큰 상·하위 15 % 일자에 게재된 주식 관련 기사를 대상으로 감성어휘를 추출하여 감성 사전을 구축하여 인터넷 뉴스 매체의 주가 상승 예측정확도를 측정하는 연구를 실행하였으나 주식 도메인에 특정되었다[4].

국방 기사 분석에 대한 연구로 Choi et. al.은 국방 기사 데이터를 처리, 분석하여 개별 문서에 대한 구조화된 정보를 추출하여 데이터베이스화 한 후 사용자가 질의하는 질문에 대해 답변할 수 있는 맞춤형 정보 분석 시스템을 제안하고 이에 필요한 기술들을 소개하였으나 데이터 수집 과정에 대한 설명이 부족하고, 시스템 안에 자연어 처리 또한 영어를 바탕으로 되어 있다.[5]

따라서 본 연구에서는 국방 관련 기사를 자동으로 수집하여 주제를 분류하고 본문을 요약하며, 기본적인 감성 분석까지 가능한 국방 기사 자동 분석 시스템 구축 방안에 대해 제안하려고 한다.

본 연구의 구성은 다음과 같다. 2장에서는 본 연구를 위한 이론적 배경에 대해 언급하고 3장에서는 시스템 기본 절차를, 4장은 구축한 시스템을 바탕으로 실험한 결과를 보여준 후 5장에서 결론을 제시한다.

2. 이론적 배경

국방 기사 자동 분석 시스템을 구축하기 위해 먼저 뉴스가 어떻게 현상을 바라보고 표현하는지를 이해하는 뉴스 프레이밍 기법을 설명하고, 이후 기사를 처리, 분석하기 위한 텍스트 마이닝 분야를 소개하고 추가적으로 기사 주제 분포 파악을 위한 토픽 모델링, 긍·부정 판단을 위한 감성 분석 기법, 사용자가 기사 내용을 빠르게 파악하기 위한 다중 문서 요약 등에 대해 언급한다.

2.1 뉴스 프레이밍(News Framing)

하나의 사실을 가지고 기사를 작성할 때 이를 규정하는 프레이밍 이론은 학회에서 중요하게 논의되고 있다. 프레임이란 현실을 구성하는 다양한 요소들 중에서 어떤 특정 면을 부각하고 재구성함으로써, 수용자가 해석하고 인지하는데 영향을 미치는 특정한 기준 혹은 틀로써, 뉴스 생산과정과 동시에 독자가 뉴스를 해석하는 과정의 기준과 방향성도 포함하고 있다[6]. 이 프레임을 통해 기사가 이슈를 긍정적으로 바라보는지 부정적으로 바라보는지에 대한 판단도 가능하다.

따라서, 뉴스 프레임을 분석하는 것은 파급효과와 영향력을 판단하는데 중요하다고 할 수 있는데, B. Burscher et. al.의 연구에 따르면 프레임 분석시 제목만을 사용할 경우 기사 전체를 사용했을 때보다 더욱 정교하게 프레임을 표현할 수 있다[7].

2.2 텍스트 마이닝(Text Mining)

텍스트마이닝은 다양한 문서들 사이에서 유의미한 정보를 추출하고 패턴을 파악하는 방법으로 문서 내 이벤트 등 특정한 정보 추출(Information Extraction)부터 사용자의 관심사항에 따라 해당 내용의 문서를 지속 제공해주는 토픽 추적(Topic Tracking), 주어진 문서들을 지정된 그룹에 할당해주는 분류(Classification), 유사 문서들끼리 그룹화 해주는 군집(Clustering), 긴 내용의 문서에서 핵심 내용을 추출·정리하여 제공하는 요약(Summarization) 등이 있다[8].

2.2.1 LDA(Latent Dirichlet Allocation)

토픽 모델링은 문서들 사이에서 구성된 키워드간의 관계를 통해 문서의 주제를 찾아내는 기법이다.

LDA는 토픽 모델링 알고리즘 중 하나로 각 문서는 여러 토픽들로 구성되어 있고 서로 다른 확률로 분포되어 있다는 가정 하에 결합확률분포와 조건부분포를 계산하여 토픽 별로 관련 키워드를 추출한다[9].

LDA는 수많은 연구에서 문서 내 토픽과 키워드 분포를 확인하기 위해 사용되고 있는데, 특히 한국사회

의 투자자 관계 개념 연구[10], 대학구조개혁평가 쟁점 분석[11], 사회문제 해결형 기술수요 발굴을 위한 키워드 추출 시스템 제안[12] 등 신문 기사를 데이터로 하는 연구에서 효과적인 분석 도구로 사용되고 있다.

2.2.2 감성 분석(Sentiment Analysis)

감성 분석은 문서나 문장에서 어떤 대상에 대한 감성을 찾아내는 것을 말하며, 주로 영화, 제품에 관한 리뷰, 블로그, 신문기사 등에서 활발하게 이루어져 왔다[13].

감성 분석을 하기 위해서는 구축한 감성사전에 있는 긍·부정 어휘들을 이용하여 문장에서 단어를 추출한 다음 사전에 정의된 감정값에 의해 그 결과로 감정을 분류하는 사전 기반 방식과 수집된 데이터를 분류하기 위해 알고리즘을 학습시키고, 학습에 사용된 감정 자질이 실제 분류시 가중치를 갖게 되어 문장의 감성을 분석하는 기계학습 방식이 있다[14].

본 연구에서는 군 관련 이슈에 긍·부정적인 기사였다고 판단할 수 있는 학습 데이터를 충분히 구비하지 못해 기계학습 방식을 적용하기에 제한되었고, 주식, 영화 리뷰 등 기존 감성 분석 연구와 도메인이 달라 해당 감성사전을 적용하기 어려운 점을 고려, 4,264건의 국방 기사에서 자주 언급된 명사, 동사, 형용사 각 3,000단어에서 긍·부정 감정을 나타내는 어휘를 선별하여 새로 감성사전을 구축하였다.

2.2.3 Text Summarization

주어진 문서를 요약하는 방법에는 문서에 있는 문장 중 중요 문장을 선택하여 제공하는 추출 기반(Extraction-based) 방식과 유의미한 단어로 재구성하여 제공하는 추상 기반(Abstraction-based) 방식으로 구분된다.

주어진 문서에서 중요한 문장을 찾고, 불필요한 문장을 제거하는 작업을 반복하여 본문보다 적은 수의 문장으로 충분한 정보량을 표현하는 추출 기반 다중 문서 요약 기술은 지속적으로 연구되어 왔으며, 최근, 딥러닝과 컴퓨팅 기술의 발달로 RNN과 집중 방식(Attention Mechanism)을 이용한 문서 요약 기술과 관련하여 학술적 연구가 활발하게 진행되고 있다[15].

LexRank 알고리즘은 중심성이 높은 문장들을 추출하는 방식으로 문서를 요약한다. 먼저 각 문장 단어별 TF-IDF 점수를 합산하여 문장의 중요 정도를 계산 후 코사인 유사도를 측정하여 각 문장간 유사도를 확인한다. 계산된 코사인 유사도 값에 임계(Threshold)를 설정하여 각 문장을 노드로 표현한 그래프를 만들고 총 노드 수와 인접 노드 수, 연결된 차수(Degree)를 Google 검색 엔진에서 페이지의 상대적 우선 순위 결정에 사용되는 PageRank로 계산하여 중요 문장을 최종 추출한다[16].

한국어의 경우 LexRank를 기반으로 문장간 코사인 유사도와 그래프 클러스터링을 통해 주제가 다양하고 긴 다중 문서 요약에 적합한 Lexrankr이 개발되어 있다[17].

3. 시스템 기본 구성

본 논문에서 제안하는 국방 기사 자동 분석 시스템은 인터넷 상의 국방 관련 기사를 자동으로 수집, 기사의 제목을 이용하여 토픽 모델링을 실시하여 기사들을 대표하는 주제들을 파악하고, 본문을 요약 및 감성 분석을 통해 사용자가 빠르게 정보를 판단할 수 있도록 지원하는 체계이다. 전체적인 시스템 프로세스는 Fig. 1과 같다.

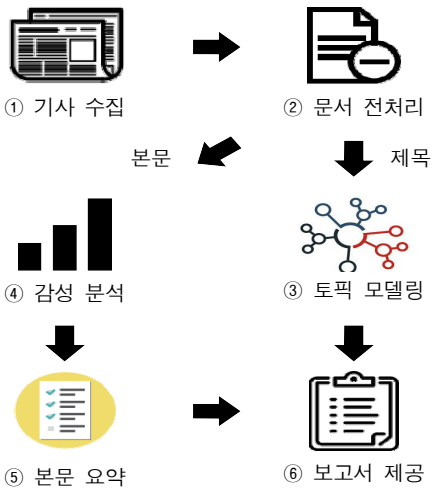

Fig. 1. System process

3.1 기사 수집

웹크롤링 기술을 이용하여 인터넷에 있는 국방 관련 기사의 제목과 본문을 자동으로 수집해주며, 키워드별, 요일별로 구분하여 수집 가능하도록 구현하였다. Python에 있는 BeautifulSoup, Request 라이브러리를 사용하였다.

3.2 문서 전처리

문서에서 불필요한 단어를 처리해주기 위한 절차로 분석에 필요한 데이터를 얻기 위해 중요하며, 본 연구에서는 분석 효율을 향상시키기 위해 명사, 동사, 형용사 품사만을 활용하였다. 도구는 Python KoNLPy 라이브러리를 사용하였다[18].

3.3 토픽 모델링

국방 관련 기사들이 어떠한 주제와 핵심 키워드들로 구성되어 있는지 분석하기 위해 전처리한 기사 제목만을 이용하여 LDA를 실시한다. Python gensim 라이브러리를 사용하였다[19].

3.4 감성 분석

기사의 감성이 긍정인지 부정인지를 구분하기 위한 절차로 '17년 전반기 국방 관련 기사 중 사드 배치, 국방 장관 후보자 내정 관련 기사 등 4,264건에서 자주 언급된 명사, 동사, 형용사 3,000 단어를 긍·부정으로 분류한 감성사전을 이용하였다.

3.5 본문 요약

시스템 사용자가 관심 있는 기사의 내용을 빠르게 이해하는 것을 지원하기 위해 기사 본문 내용을 요약하며, Python lexrankr 라이브러리를 사용하였다.

3.6 보고서 제공

데이터 분석 결과를 바탕으로 사용자의 수요에 맞는 분석 보고서를 제공하는 절차로 시스템에 의해 자동으로 작성되는 것을 지향하나 향후 연구에서 구현할 예정이다.

4. 실험 결과

이번 장에서는 실제 네이버 뉴스에서 제공한 국방/외교 범주에 있는 기사를 대상으로 자동 분석한 결과를 보여준다.

4.1 기사 수집

뉴스 데이터는 웹크롤링 기법을 이용하여 네이버 뉴스에서 제공하는 국방/외교 범주에 있는 '17. 1. 1. ~ '17. 6. 30.까지 기사 39,403건을 요일별로 수집하였으며 그 중 본문이 중복되는 포토기사를 제외한 32,465건을 데이터로 설정하였다.

4.2 문서 전처리

토픽 모델링과 감성 분석을 위해 기사 32,465건을 KoNLPy 라이브러리를 이용하여 처리한 결과 제목은 423,997개, 본문은 14,454,103개의 토큰으로 구성되어 있었으며, 그 중 2음절 이상의 명사와 동사, 형용사를 제외한 다른 품사를 제거하여 제목은 197,171개(46 %), 본문은 6,070,955개(42 %)의 토큰으로 축소되었다.

4.3 토픽 모델링

32,465건의 기사 제목만을 이용하여 LDA 기법을 통해 토픽 모델링을 실시한 결과는 Table 1과 같다. LDA 적용시 문서 토픽 분포 정도인 변수 α(설정 : 0.25)와 토픽 내 키워드 분포 정도인 β(설정 : 0.002), 샘플링 정도인 iteration(설정 : 3,000)은 연구자가 직접 설정해 준 값으로 실시하였으며, 토픽과 토픽당 키워드는 실험 결과 각 10개가 가장 가독성이 좋다고 판단하여 고정 값으로 설정하였다.

Table 1. News title topic modeling result

토픽	키워드
1	대화, 트럼프, 동맹, 특사, 대통령, 협력, 한미, 비핵화, 환영, 도착
2	장관, 한민구, 외교, 총리, 인사, 국방, 회담, 개최, 임관, 국방부
3	정상, 위안부, 합의, 일본, 대통령, 문제, 외교부, 정부, 주한, 협상
4	무인기, 조사, 행사, 전략, 개혁, 필요, 사드, 후보, 전문가, 민간
5	발언, 외교부 장관, 준비, 철회, 주민, 보고서, 이동, 평화, 기여, 공개
6	발사, 미사일, 부통령, 대통령, 시험, 내년, 북한, 성공, 탄도미사일, 규탄
7	악수, 군사, 사령관, 대표, 방문, 합참, 연합, 대통령, 비용, 사드
8	강경화, 대통령, 사드, 만나, 방미, 배치, 문재인, 시작, 반대, 의원
9	제재, 대북, 대통령, 논란, 무역, 해결, 파견, 압박, 착수, 방명록
10	성주, 장병, 해군, 기념식, 면담, 참석, 국민의당, 일정, 관리, 정치

토픽1은 한미정상회담, 토픽2는 총리·장관 주관 회담, 토픽3은 위안부 합의 문제, 토픽4는 북 무인기 이슈, 토픽5는 사드 철회에 대한 이슈, 토픽6은 펜스 미 부통령 방한 및 북 탄도미사일 규탄, 토픽7은 대통령 취임 후 합참 및 연합사 순시, 토픽8은 사드 배치 반대 이슈, 토픽9는 북핵 대응 관련 내용임을 알 수 있다. 토픽10은 사드 배치에 대한 국회 반응과 서해 수호의 날 기념식 관련 키워드가 추출된 것으로 판단된다.

토픽 모델링 결과를 통해 2017년 전반기 동안 국방/외교 분야에서 어떤 주제를 가지고 기사가 보도 되었는지 확인 가능하다.

4.4 감성 분석

'17년 전반기 국방 관련 기사 중 사드, 장관 후보자 관련 기사 등 4,264건에서 자주 사용된 2음절 이상의 명사, 동사, 형용사 각 3,000단어를 추출하여 '강력하다, 좋다, 최고' 등 긍정적 단어와 '심각하다, 위험하다, 위협' 등 부정적 단어를 구분하여 사전을 구축하였다(긍정 498단어, 부정 634단어).

Table 2. News sentiment analysis result(example)

기사(보도 일자)	점수	결과
해군, 실수로 동해에 미사일 투하(1. 1.)	18.18	부정
미국·이란 핵합의 파기 가능성과 북한의 배신 트라우마(2. 8.)	53.13	중립
北, 김정일 사망 직후 軍 간부들 표정까지 사찰(2.22.)	7.69	부정
'성매매 혐의' 육사생도 3명, 졸업 하루 앞두고 퇴교 조치, 익명 제보 조사는 내규 위반 논란도(2.23.)	33.33	부정
정부, 전 세계 대사관에 재외국민 보호 강화 지시(2.23.)	68.75	긍정
칼빈슨 항공모함 동해서 北 미사일 격추 훈련 시작(4.29.)	66.67	긍정
미국내 최대 친한그룹 '주한미군전우회' 출범(5. 2.)	94.59	긍정
文측, 황기철 전 해군참모총장 영입…"준비된 후보 증명"(5. 3.)	47.05	중립
文-트럼프 백악관서 만났다…상견례 시작으로 환영 만찬 등 일정 소화(6.30.)	95.45	긍정

감성사전을 기반으로 기사 본문에서 긍정 단어가 나온 횟수와 부정 단어가 나온 횟수를 비교하여 문서의 감성 점수를 계산하였으며 식은 (1)과 같다.

$$score(i) = \frac{\sum hit(p_i)}{\sum hit(p_i) + \sum hit(n_i)} \times 100 \quad (1)$$

score(i), hit(p_i), hit(n_i)는 각각 i번째 기사의 감성 점수, 긍정 단어 출현 수, 부정 단어 출현 수이고, 기사의 긍·부정 점수는 긍·부정 단어가 출현한 전체 수에서 긍정 단어가 차지하는 비율로 계산하였으며, 분모가 0이 되는 것을 방지하기 위해 초기 값으로 hit(p_i)와 hit(n_i) 모두 1을 설정하였다. 40점 미만은 부정, 40점 이상 60점 미만은 중립, 60점 이상은 긍정으로 분류하였으며 결과는 Table 2 및 Table 3과 같다.

Table 3. News sentiment analysis result(count)

긍정	중립	부정	계
12,710건	11,511건	8,244건	32,465건

감성 분석은 감성사전 단어의 출현 횟수로 비교하였기 때문에 기사의 논조가 긍정인지 부정인지를 정확하게 판별하기에 제한이 따른다. 실제로 2017년 1월 5일 뉴시스에서 보도한 "정부 北인권 실태조사 착수... 가해자 몽타주 작성" 제하 기사의 경우 북한인권 실태조사에 대한 내용이지만 기사 본문에 포함된 단어 중 상당수가 '제거, 반발, 겨냥, 침해, 가해자' 등의 부정 감성사전에 등록된 단어여서 부정 기사로 분류되었다. 하지만, 실무적인 차원에서 1차 분류 용도로 사용 가능할 것으로 판단된다.

4.5 본문 요약

본문 요약 단계에서는 수집된 뉴스 본문을 대상으로 중요 내용을 표현할 수 있는 최소의 단락으로 요약하여 제공하여 준다.

본 시스템에서 사용하는 문서요약은 추출 기반(Extraction-based) 알고리즘으로 실제 기사에서 사용한 문장 중 가장 중심 점수가 높은 문장을 선택하기 때문에 어색하지 않고 문장 구조가 완전한 요약문을 제공한다.

Table 4는 Lexrankr 결과로 생성한 기사 본문의 요약문 예시이다.

Table 4. News article summarization result(example)

기사		내용
1	원문	통일부는 1일 북한 김정은 국무위원장의 신년사가 정책 비전 제시에 있어 한계를 드러냈다고 평가했다. '대남(對南)', '대미(對美)' 관계 등 대외 정책에서는 핵 무력을 중심으로 한 공세적 태도를 예고했다는 분석이다. (중략) 통일부는 "대통령 실명을 언급하고 '반(反) 통일 세력 분쇄'를 주장한 것은 현재 남측 내부 정세가 자신들(북측)에게 유리하게 전개됐다는 판단에서 비롯된 것"이라며 "향후 통전 차원의 평화 공세를 강화할 것으로 예상된다"고 밝혔다. 아울러 "고위급 접촉 등 당국 간 회담에 대해 별도의 언급을 하지 않은 만큼, 올해도 민간 차원의 접촉을 통한 대남 흔들기를 강화할 것으로 전망된다"고 덧붙였다.
	요약	통일부는 1일 북한 김정은 국무위원장의 신년사가 정책 비전 제시에 있어 한계를 드러냈다고 평가했다. 통일부는 "이례적으로 김정은이 '심부름꾼', '충복'임을 강조한 것은 성과 부진에 대한 비난을 회피하고, '인민 중시'를 김정은 시대 브랜드로 만들어 대중적 기반을 구축하려는 의도"라며 "세도, 관료주의, 부정부패 극복 및 '인민에의 멸사복무'를 요구하며 총동원 체제를 위한 사회통제 강화 의지도 시사했다"고 평가했다
2	원문	한미 양국의 산업 담당 장관이 직접 만나 한미 자유무역협정(FTA)을 "양국 간 경제협력의 기본틀"이라고 평가했다. 도널드 트럼프 미국 대통령 취임 이후 한미 FTA를 둘러싼 미국 내 부정적 반응이 커지는 가운데 나온 긍정적인 인식의 공유라는 점에서 의미가 크다는 분석이 나온다. 주형환 산업통상자원부 장관은 8일(현지 시간) 미국 워싱턴에서 윌버 로스 미국 상무장관과 회담을 갖고 이 같은 내용을 골자로 한 통상·산업협력 강화를 논의했다고 밝혔다. (중략) 한미 안보·경제 동맹 강화라는 큰 틀의 협력 체계와는 별개로 미국의 이해관계에 어긋나는 부분에 대해서는 강한 압박을 가해 '미국 우선주의' 기조를 이어갈 것이라는 전망이다.
	요약	산업부는 "한미 FTA가 양국 경제의 기본 틀로서 FTA 발효 후 5년간 양국 간 교역 및 투자 확대에 기여해 온 객관적 성과를 미국과 공유했다"고 밝혔다. 정부 안팎에서는 한미 FTA에 대한 미국 정부의 호의적 반응이 중국의 사드(THAAD·고고도미사일방어체계) 경제 보복에 따른 안보동맹 강화 차원에서 나온 게 아니냐는 분석을 하고 있다
3	원문	이순진 합동참모본부 의장은 1일 새해를 맞아 '하늘의 지휘소'라 불리는 피스아이(항공통제기)에 탑승, 군사대비태세를 점검했다고 합참이 밝혔다. 이 의장의 피스아이 작전지휘비행은 장병들의 노고를 격려하고 새해에도 변함없이 확고한 군사대비태세 유지를 당부하기 위해 실시됐다고 합참은 설명했다. (중략) 이 의장은 서해-중부전선-동해를 아우르는 전술비행 과정에서 해병 6여단장, 육군 6사단 연대장, 해군 양만춘함장, 공군 F-16편대장 등과 격려통화를 실시했다고 합참은 설명했다. 한편 이 의장이 이날 탑승한 피스아이 E-737은 '하늘의 지휘소'라 불리는 항공통제기로 기체에 공중감시레이더를 장착, 공중에서 조기경보·항공기 통제·전장 관리 등 다양한 임무를 수행한다. 공군은 2011년 피스아이 1호기를 도입한 이후 현재 총 4대를 도입·운용 중에 있다.
	요약	이 의장은 피스아이의 탑승한 채로 백령도의 해병6여단장, 중부전선의 일반전초(GOP)연대장, 동해상에서 작전중인 양만춘함 함장, 비상출격한 전투기 조종사로부터 방어태세에 대한 보고를 받았다. 이 의장은 서해-중부전선-동해를 아우르는 전술비행 과정에서 해병 6여단장, 육군 6사단 연대장, 해군 양만춘함장, 공군 F-16편대장 등과 격려통화를 실시했다고 합참은 설명했다

Table 5. News article summarization result(rate)

본문 토큰	요약 토큰	요약률
14,454,103	5,247,280	64 %

Table 5는 본문과 요약문 간에 토큰 수를 비교한 요약률을 나타낸 것으로 계산식은 (2)와 같다.

$$\text{Rate} = 1 - \frac{\text{Summarization Token}}{\text{Total Text}} \quad (2)$$

Summarization은 요약된 토큰의 수, Total Text는 전체 본문의 토큰 수이다.

5. 결 론

본 연구를 통해서 우리는 국방 관련 기사를 자동으로 수집하여 분석할 수 있는 시스템 구축이 가능함을 확인하였다.

본 연구의 의의는 ① 텍스트 마이닝의 도메인으로 국방 관련 기사를 사용함으로써 연구 영역을 확장하였고, ② 국방 기사에 적합한 감성사전을 구축하여 기본적인 감성 분석이 가능함을 확인하였으며, ③ 외국 기사가 아닌 국내 기사에 중점을 두어 활용성을 높였다.

본 시스템 구축시 군이 가질 수 있는 이점은 ① 온라인상에 게시된 수많은 군 관련 기사들을 확인하고 주제 분류 및 요약하며, 1차원적인 감성 분석까지 수행 가능하고, ② 사용자에게 근실시간으로 군에 대한 언론 보도 내용을 제공하여 빠른 결심을 보좌하며, ③ 기사 수집에서 분석까지 자동화하여 적은 인력으로도 높은 효율을 얻을 수 있다.

하지만 본 연구의 한계점도 분명히 존재한다. 먼저, 본 연구는 네이버에서 제공하는 인터넷 뉴스를 기반으로 수집하였기 때문에 트위터 등 SNS에 게시되는 기사까지 수집·분석 가능한 시스템 구축에 대해서는 추가적인 연구가 필요하다. 또한, 감성 분석의 경우 어휘 빈도수 기반의 기본적인 분석으로 정확한 기사 논조 분석이 어려워 사람을 통한 재분류 작업이 필요하다. 향후 기계학습을 통한 감성 분석을 할 수 있는 추가적인 연구가 필요할 것으로 판단된다. 마지막으로 ROUGE[20] 점수(Recall-Oriented Understudy for Gisting Evaluation Score) 기반의 본문 요약의 정확도 측정을 위한 추가적인 검증용 데이터 확보가 필요하다.

References

[1] Amy E. Jasperson, Dhavan V. Shah, Mark Watts, Ronald J. Faber & David P. Fan, “Framing and the Public Agenda: Media Effects onthe Importance of the Federal Budget Deficit,” Political Communication, Vol. 15, pp. 205-224, 1998.

[2] Bjorn Burscher, Rens Vliegenthart, and Claes H. de Vreese, “Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue,” Social Science Computer Review, Vol. 34, No. 5, pp. 530-545, 2016.

[3] M. Gam, et. al., “A Study on Differences of Contents and Tones of Arguments among Newspapers Using Text Mining Analysis,” Journal of Intelligence and Information Systems, Vol. 18, No. 3, pp. 53-77, 2012.

[4] S. Cheon, et. al., “A Comparative Study on the Accuracy of Stock Price Prediction by Medium by Opinion Mining of News Contents,” Korea Intelligent Information System Society Spring Conference, pp. 133-137, June, 2013.

[5] J. Choi, et. al., “Customized Information Analysis System Using National Defense News Data,” The Journal of the Korea Contents Association, pp. 457-465, Vol. 10, No. 12, 2010.

[6] M. Im, C. An, G. Gam, H. Yu, “A Study of the News Frame in Newspapers : Frame Analysis of Park Geun-hye in Chosunilbo and Hankyereh,” Journal of Communication Science, Vol. 10, No. 3, pp. 457-498, 2010.

[7] Bjorn Burscher, Rens Vliegenthart, and Claes H. de Vreese, “Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue,” Social Science Computer Review, Vol. 34, No. 5, pp. 530-545, 2016.

[8] R. Balamurugan, “A Review on Various Text Ming Techniques and Algorithms,” International Conference on Recent Innovations in Science, Engineering and Management, pp. 837-848, November, 2015.

[9] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation,” Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[10] A. Mun, “Investor Relations Concept in the Korea Society : 1994~2014 Newspaper Content Analysis and Semantic Network Analysis,” Journal of Public Relations, Vol. 20, No. 1, pp. 50-78, 2016.

[11] J. Kim, “Analysis of Issues on the College and University Structural Reform Evaluation Using Text Big Data Analytics,” Asian Journal of Education, Vol. 17, No. 1, pp. 409-436, 2016.

[12] D. Jeong, J. Kim, G. Kim, J. Heo, B. On, M. Kang, “A Proposal of a Keyword Extraction System for Detecting Social Issues,” Journal of Intelligence and Information Systems, Vol. 19, No. 3, pp. 1-23, 2013.

[13] Y. Kim, et. al., “Korean Text Emotion Classification Using Machine Learning,” Korea Entertainment Industry Association 2013 Fall Conference Session Ⅱ, Vol. 13, pp. 206-210, November, 2013.

[14] J. Lim, J. Kim, “An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter,” Journal of Korea Multimedia Society, Vol. 17, No. 2, pp. 232-239, 2014.

[15] R. Nallapati et. al., “SEQUENCE-TO-SEQUENCE RNNs for Text Summarization,” Workshop Track-International Conference on Learning Representations, pp. 1-4, May, 2016.

[16] G. Erkan et al, “LexRank: Graph-based Lexical Centrality as Salience in Text Summarization,” Journal of Artificial Intelligence Research, Vol. 22, pp. 457-479, 2004.

[17] J. Seol, S. Lee, “Lexrankr: LexRank based Korean Multi-Document Summarization,” Communications of the Korean Institute of Information Scientists and Engineers 2016 Winter Conference Language Engineering Session, pp. 458-460, December, 2016.

[18] E. Park, S. Cho, “KoNLPy: Korean Natural Language Processing in Python,” The 26th Annual Conference on Human and Language Technology, pp. 133-136, October, 2014.

[19] R. Rehurek, P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45-50, May, 2010.

[20] Chin-Yew Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74-81, July, 2004.