ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2021 September: 27(3): 75~94 http://dx.doi.org/10.13088/jiis.2021.27.3.075

텍스트 마이닝을 이용한 공군 부사관 지원자 자기소개서의 차별적 특성 분석

권 혁
연세대학교 산업공학과
(gwonkh@naver.com)

김우주
연세대학교 산업공학과
(wkim@yonsei.ac.kr)

저출산 문제로 인한 병역자원 감소와 병 복무기간 단축에 따른 군 간부 대비 병 복무 선호 현상은 우수한 군 간부확보 정책에 대한 추가적인 고찰을 필요로 한다. 이와 관련된 연구들은 대부분 사회과학에서 주로 사용되는 방법론으로 분석하였으나, 본 연구는 대량의 문헌조사에 적합한 텍스트 마이닝의 방법론으로 접근한다. 이를 위해, 본 연구는 공군 부사관 지원자 자기소개서에서 차별적인 특성의 단어들을 추출하고 합격 및 불합격의 극성을 분석한다. 본 연구는 총 3단계로 이루어졌다. 첫번째, 지원분야를 일반분야와 기술분야로 나누고, 자기소개서에서 특성을 가지는 단어들을 분야별 빈도수 비율의 차이대로 순서화 한다. 각 지원분야별 비율의 차이가 클수록 해당 지원분야의 특성을 나타내는 것으로 정의하였다. 두번째, 이 특성을 나타내는 단어들을 LDA를 통해 단어들의 Topic을 군집화하고 이를 바탕으로 Label을 정의하였다. 세번째, 이 군집화 된 지원분야별 단어들을 L-LDA를 통해 합격과 불합격의 극성을 분석하였다. L-LDA값의 차이가 합격에 가까울수록 합격자들이 많이 사용하는 단어로 정의하였다. 본 연구를 통해, 공군 부사관 자기소개서의 차별적 특성을 추출하기에는 LDA보다 L-LDA가 더 적합함을 알 수 있다. 또한, 이러한 방법론은 별도의 서면 또는 대면 설문 방식이 아니라, 대량 문서에 대한 텍스트 마이닝 기법을 적용하여 분석시간을 단축하고, 전체 모집단에 대한 신뢰성을 높일 수 있다. 따라서 본 연구인 공군 부사관 선발결과 분석을 통해, 선발제도 및 홍보제도에 활용 가능한 정보를 제공하고, 군 인력획득 분야 연구에 있어 활용 가능한 방법론을 제안하고자 한다.

주제어 : 공군 부사관, 자기소개서, 텍스트 마이닝, LDA(latent Dirichlet allocation), L-LDA(Labeled latent Dirichlet allocation)

논문접수일 : 2021년 5월 21일 논문수정일 : 2021년 7월 23일 게재확정일 : 2021년 7월 29일
원고유형 : 학술대회 Fast-Track 교신저자 : 김우주

1. 서론

저출산 문제와 병 복무기간 단축은 우수한 군 간부 선발에 대한 고민을 갖게 하고 있다. 우리나라는 1984년 저출산 사회 진입, 2017년 생산인구가 감소 시작, 2018년 고령사회로 진입했고, 2025년에는 초고령사회에 진입할 것으로 전망된다(Oh et al., 2020). 이와 더불어, 최첨단 무기전쟁인 미래전 환경에 대비하여 국방인력을 감축하고, 상비병력은 현역병 중심에서 숙련 간부 중심으로 전환이 진행 중이다. 또한, 병력 중심의 군을 최첨단 전력 중심으로 정예화하고, 청년들의 병역 부담 완화 및 조기 사회진출을 위한 병 복무기간 단축이 2018년 시행되었다. 따라서 병역자원 감소와 의무복무로서 군 간부보다 병 복무를 선호하는 현상이 관측되고 있다.

이러한 상황을 극복하기 위해 우수한 군 간부 확보에 관한 연구가 최근 많이 진행되고 있다.

주요 연구로는 저출산 상황에서 우리 군에 대한 시사점을 찾기 위한 외국군 인력획득정책 연구(Dohkgoh et al., 2017)와 직업군인을 선택하는 영향요인과 직업 선택 만족도와의 관련성에 관한 연구(Baek et al., 2019) 등이 있다. 하지만, 대부분의 연구가 사회과학에서 주로 사용되는 방법론으로 분석되어 데이터 자체의 의미를 활용하기에는 어려움이 있음에 따라, 본 연구는 자연어로 구성된 텍스트 데이터에서 패턴을 추출하는 텍스트 마이닝 방법론을 활용하였다.

기존에도 국방분야에서 텍스트 마이닝에 관한 연구가 있었으나, 주로 기술, 정보, 공보, 행정, 병영, 학술 분야 등에 한정적이었다. 주요 연구로는 장갑전투차량의 기술변화 추세와 미래 개발 이슈(Jeon et al., 2020), 북한 보도내용과 북한 도발과의 연관성 분석(Lee et al., 2016), 국방 관련 기사에 관한 주제를 분류 및 요약하고 감성분석 시스템 구축에 관한 연구(Kim et al., 2018), 군 조직의 온나라시스템을 바탕으로 정량적 업무분석에 관한 연구(Lim et al., 2019), 병사의 생활지도기록부 및 SNS를 활용한 사고예측 모델 개발에 관한 연구(Yoon et al., 2015), 군사학 분야 학술 논문의 세부 학문분류에 관한 연구(Bae et al., 2020) 등이 있다. 하지만, 민감한 개인정보 텍스트로 구성된 대량의 데이터가 구축되어 있는 군 인력획득 분야에서는 텍스트 마이닝 방법론을 적용한 연구가 없었다.

따라서 본 연구는 개인정보 비식별 조치한 2019년도 공군 부사관후보생 지원자들의 자기소개서를 바탕으로 군 간부 지원자들의 합격 및 불합격 성향을 텍스트 마이닝으로 분석한다. 이를 통해 선발제도, 홍보제도 등 인력획득정책에 활용 가능한 정보를 제공하고, 군 인력획득분야를 포함한 군 인사분야 연구에 있어 활용 가능한 방법론을 제안하고자 한다. 예를 들면, 특정 단어가 합격의 극성을 높게 나타내는 단어로 나타났을 때, 이 단어가 해당 분야의 직무수행에 있어 중요한 요소인지를 검토하고, 필요하다면 선발정책을 개선할 수 있다. 그리고 직무에 필요한 특정 단어의 성향을 가진 지원 대상자들에게 충분히 홍보가 되고 있는지에 대하여 검토하고, 부족하다면 홍보정책을 개선할 수 있다. 또한, 본 연구의 방법론이 다른 간부 모집전형, 진급심사, 장기복무 선발심사, 보직심의 등 다른 형태의 군 인사분야 대량 문서에도 적용 가능한 방법론임을 제안하고자 한다. 그리고 본 연구는 L-LDA가 LDA보다 차별적 특성을 나타내는 단어의 극성을 분리하는데 있어서 더 적합함을 제안하고자 한다.

2. 관련연구

2.1. 자기소개서 관련연구

자기소개서는 지원자가 인사(선발)담당자에게 자신이 지원 직무에 최적의 인재라는 사실을 설득하는 글이다(Shin, 2015). 우수한 자기소개서는 핵심이 간결하고, 기업(학교)의 직무(전공) 이해도가 높으며, 자신의 역량을 충분히 표현하는 유형이다. 또한, 궁극적인 목적은 기업(학교)의 인재상에 얼마나 적합한지를 진정성 있는 자신만의 스토리를 구조화하여 설득하는 것이다(Kim, 2014). 따라서, 자기소개서에는 지원자들의 성장과정, 성격의 장단점, 지원동기, 입사(입학) 후 포부 등을 알 수 있다. 이를 잘 활용하면 어떤 성향의 지원자들이 지원을 하는지와 어떤 성향의 지원자들을 선발하고 있는지를 분석할 수 있다.

최근, 자기소개서와 관련된 텍스트 마이닝을 활용한 연구들이 다수 진행되고 있다.

Lee et al.(2018)은 TF-IDF와 LDA를 활용하여 자기소개서를 객관적이면서 정량적으로 평가하기 위한 새로운 방법론들을 제안하였다. 토픽모델링을 통한 지원대학별, 출신고교별, 문항별 데이터를 구성하여 유사도를 다양한 차원에서 포착할 수 있도록 하였다. 하지만, 주어진 토픽들을 가지고 지원자, 합격자, 불합격자의 성향을 분석하기는 어려웠다.

Kim et al.(2020)은 자기소개서를 대상으로 Doc2Vec을 활용하여 문서를 합격, 불합격문서로 이진분류 하는 모델을 연구하였다. 이를 통해 기업 인사담당자들은 채용과정에서 소모되는 비용 및 시간을 절감하고 지원자들은 제출 전에 평가를 미리 받아 보완할 수 있도록 하는 것이 연구의 목적이었다. 그러나 이 연구모델은 단순히 합격자와 불합격자의 분류에만 활용하였다.

2.2. 국방분야 텍스트 마이닝

국방분야는 많은 텍스트들을 보유함에도 불구하고, 군사 보안의 문제로 텍스트 마이닝 연구가 민간과 같이 활발하지 않다. 따라서 공개되어 있는 외부의 텍스트를 바탕으로 국방분야 내 소수의 분야에서만 텍스트 마이닝을 이용한 연구가 진행되고 있다.

기술분야에서는 LDA를 이용하여 장갑차의 기술동향 및 미래 개발이슈를 탐색했고(Jeon et al., 2020), 정보분야에서는 TF-IDF와 감성분석을 이용하여 북한의 보도내용과 관련과 북한 도발의 연관성 분석이 있었다(Lee et al., 2016). 공보분야에서는 LDA, 감성분석, lexrank를 이용한 국방 관련 기사의 주제 분류, 감성분석, 요약에 관한 연구가 있었고(Kim et al., 2018), 행정분야에서는 TF-IDF, 감성분석을 이용한 온나라시스템과 정량적 업무분석에 관한 연구가 있었다(Lim et al., 2019). 병영분야에서는 TF-IDF와 LDA를 이용한 사고예측모델 개발에 관한 연구가 있었고(Yoon et al., 2015), 학술분야에서는 LDA와 토픽네트워크를 이용한 군사학 분야 논문의 세부 학문 분류에 관한 연구가 있었다(Bae et al., 2020).

관련 연구들을 확인한 결과 텍스트 마이닝을 이용한 군 인력획득 분야의 연구는 없었고, 이용된 방법론에서도 L-LDA 방법론을 이용한 연구는 없었다.

2.3. Text Mining

텍스트 마이닝은 대량의 비정형 문서에서 의미 있는 패턴의 정보를 정제하여 추론하는 과정이다(Tan, 1999). 구체적인 텍스트 마이닝의 방법론에는 비구조화된 데이터 집합에서 의미 있는 문서를 쉽게 찾는 Information Retrieval, 컴퓨터 또는 AI가 인간의 언어를 분석, 이해, 생성하도록 하는 Natural Language Processing, 비구조화된 문서에서 의미있는 정보를 추출하는 Information Extraction from text, 대량의 문서 또는 장문의 문서를 간결하게 요약하는 Text Summarization, 라벨이 없는 문서 집합을 군집화하거나 숨겨진 구조를 찾는 Unsupervised Learning Methods, 문서 분류를 학습한 모델을 바탕으로 새로운 문서의 분류를 예측하는 Supervised Learning Methods, 확률기반의 토픽모델과 같은 Probabilistic Methods for Text Mining, 지속적으로 생산되는 뉴스 및 SNS에서 텍스트 마이닝 하는 Text Streams and Social Media Mining, 온라인에서 제품 등의 리뷰

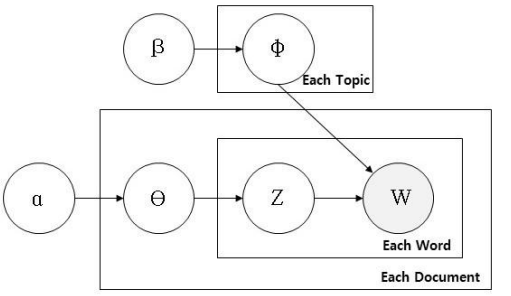
나 사용자 의견에 대한 중요한 정보를 찾는 Opinion Mining and Sentiment Analysis, 방대한 양의 의학텍스트에서 지식을 효과적으로 추출하는 Biomedical Text Mining이 있다(Allahyari et al., 2017).

본 연구에서 이용한 Probabilistic Methods for Text Mining은 대표적으로 LDA와 같은 토픽 모델링이 있으며, 이는 다양한 분야에서 적용되었다. 최근의 연구분야로, 문학 분야에서는 LDA와 네트워크분석 등을 이용하여 고전 추리소설의 두 작가 간 문체적 차이를 연구하였고(Moon et al., 2019), 언론 분야에서는 LDA와 가우시안 스무딩 기법을 이용하여 온라인 뉴스에서 사건별 네트워크를 구축하여 요약적 사건정보를 제공하는 기법에 대해 연구하였으며(Lee et al., 2018), 전자 상거래분야에서는 LDA와 k-NN 기법을 이용하여 온라인 리뷰를 바탕으로 상품의 평가기준에 대하여 연구하였다(Lee et al., 2020). 이와 다르게, 본 연구는 공군 부사관 후보생 자기소개서의 특성에 맞게 LDA와 L-LDA를 이용하였다.

2.3.1. LDA

LDA는 대량의 문서에서 개별 문서의 토픽(주제)을 확률적으로 찾아 줄 수 있는 방법론이다. 즉, 특정 문서 내 각 단어가 어떤 토픽에서 어떤 확률 값으로 발현되었는지를 추정할 수 있는 생성방식의 모델이다(Blei et al, 2003). 즉, LDA는 특정 문서에서 여러 토픽의 분포를 추정하고, 전체문서에서 토픽의 분포를 추정하며, 특정한 단어가 어느 토픽에 포함되는지의 확률도 추정할 수 있다(Teh et al., 2006).

LDA는 <Figure 1>과 같다. 각각의 토픽은 β라는 파라미터의 디리클레 분포를 따르는 토픽



〈Figure 1〉 LDA structure

별 단어분포 ϕ로 나타난다. 그리고 각각의 문서는 α라는 파라미터의 디리클레 분포를 따르는 문서의 토픽분포 θ로 나타난다. 그리고 각각의 단어에서는 토픽분포 θ에서 토픽 Z를 추정하고, 토픽별 단어분포인 ϕ와 Z에서 단어 W를 추정한다.

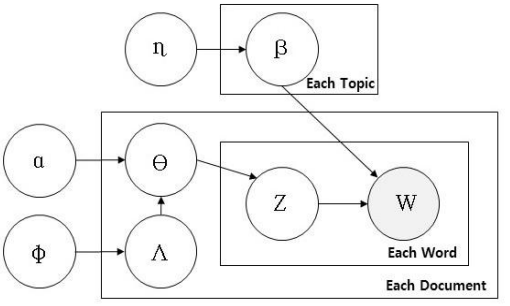
또한, 적절한 토픽 수를 확인하기 위해 Perplexity를 활용한다(Blei et al, 2003). 이 수치가 낮을수록 특정 토픽 수의 확률 모델이 성능이 높다는 것을 의미한다. 또한, 이를 보완할 수 있는 Coherence 수치를 적용하여 특정 토픽 수의 확률 모델이 상위 단어들이 얼마나 높은 유사도를 갖추었는지를 평가할 수 있다(Newman, 2010).

적절한 토픽 수를 바탕으로 주어지는 토픽과 그 토픽에 해당하는 단어들을 나타낼 수 있다. 그리고 이 토픽들에 대한 라벨링을 하는 방법론은 다양하게 있다. 주된 방법론들 중 하나는 객관화된 지표 또는 인덱스들을 활용하여 해당 토픽의 단어들이 어떤 지표 또는 인덱스에 가까운지 매칭하여 선택하는 방법이 있다(Kim et al., 2016).

2.3.2 L-LDA

LDA가 대량의 문서에서 비지도학습으로 특

정 토픽 수를 찾아준다면, L-LDA는 지도학습으로 태그 또는 평점과 같은 여러 개의 주제가 라벨링 되어있는 집합으로 제한하고 라벨링별 토픽 단어들을 분석한다. 라벨을 합격과 불합격으로 가정하면, 단어들이 합격에 가까운지 불합격에 가까운지에 대해 L-LDA 확률 값으로 극성을 분석할 수 있다.



〈Figure 2〉 L-LDA structure

L-LDA는 <Figure 2>와 같다. 각각의 토픽은 η라는 파라미터의 디리클레 분포를 따르는 토픽별 단어분포 β로 나타난다. 그리고 ϕ라는 베르누이 분포를 따르는 Λ가 라벨이 포함되는 여부인 0 또는 1의 값을 가지는 벡터 값으로 나타난다. 그리고 그 문서는 α Λ라는 파라미터의 디리클레 분포를 따르는 문서의 토픽분포 θ로 나타난다. 그리고 각각의 단어에서는 토픽분포 θ에서 토픽 Z를 추정하고, 토픽별 단어분포인 β와 Z에서 단어 W를 추정한다. LDA는 해당 단어가 어떤 토픽에서 발현하는지를 분석할 수 있다면, L-LDA는 문서에 라벨을 통하여 평점 또는 이진분류의 라벨에서 단어의 분포 및 발생 확률을 추정할 수 있다. L-LDA는 라벨에 따른 토픽의 차이를 보여줄 수 있다는 장점이 있고, LDA보다 높은 성능을 보여준다(Ramage et al., 2009).

3. 연구방법

3.1 데이터

3.1.1 데이터 형태

실험에 사용한 데이터는 개인정보 비식별화한 2019년도 공군 부사관후보생 지원자 3개 기수 총 6,153명의 자기소개서이다. 자기소개서는 총 4개의 문항으로 구성되어 있다. 문항은 가정 및 성장환경, 성장과정, 자아표현, 지원동기이다. 그리고 각 자기소개서별로 기수, 성별, 지원분야, 전형단계별 합격여부의 라벨로 구성되어 있다. 모집절차는 <Table 1>과 같다. 본 연구에서는 공군 부사관후보생의 합격 및 불합격 기준을 3차 전형으로 정의한다. 이후의 입영전형에서는 지원자가 합격자로서 훈련부대에 입영하기 때문에, 3차 전형은 선발전형에서 최종 단계이다.

〈Table 1〉 Recruitment process

1 차 전형	2 차 전형	3 차 전형	입영 전형	군사 훈련	
필기 시험 (200 점) * 가점 : 자격증 등	신체 검사 /면접 (25 점)	최종 선발 위원회 (적/부)	신체/ 인성 검사 (적/부)	11 주	**임관**

공군 부사관후보생 지원자 자기소개서는 선발전형에서 직접적인 평가요소에 포함되지 않는다. 그러나 자기소개서가 이용되는 2차 전형 면접은 지원자 대상 공군 부사관의 군인상(軍人像)과 적합도를 점수화하고, 결격자들을 판단하기 때문에 선발단계에서 중요한 평가요소이다. 이처럼 자기소개서는 면접의 기초자료로서 평가요소에 중요한 영향을 미치기 때문에, 지원자들은

자기소개서에 자신의 성향을 최대한 반영하고자 한다. 그러므로 지원자들의 자기소개서는 지원자들의 성향을 분석할 수 있는 중요한 데이터이다.

3.1.2 데이터 전처리

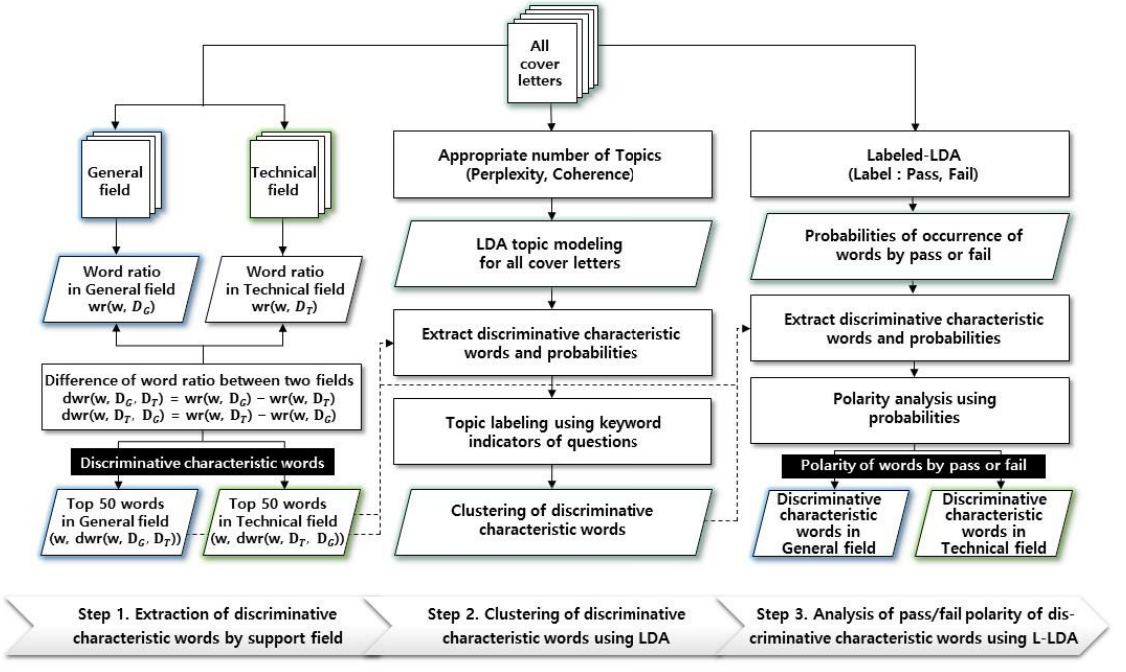
실제 자기소개서의 내용을 분석한 결과, 성격을 성장환경 또는 자아표현에 작성하는 등의 지원자 간의 문항별 자기소개서 작성의 불일치들이 발견되었다. 이를 개선하기 위해 전처리 과정에서 4개 항목의 자기소개서에 각 지원자별 정보를 동일하게 부여하여, 지원자 당 4개의 행 정보로 나타냈고, 각각의 행은 1개 자기소개서 항목만 가진다.

그리고 Python의 한국어 자연어처리 패키지인 KoNLPy 내 Twitter모듈을 이용하여 명사들을 추출하였다. 그 과정에서 불용어사전으로 이, 그, 게다가, 모든 등과 같은 의미 없는 단어 명사들 976개를 제거하였으며, 인코딩 오류가 발생한 자기소개서 222개를 제외하였다. 최종적으로 3차 시험 합격자를 최종 합격자로 정의한 24,390개의 자기소개서에 대해 합격과 불합격 라벨을 실험을 진행하였다.

3.2. 연구절차

전처리한 자기소개서 데이터를 바탕으로, 본 연구절차의 전체 개요는 다음과 같다.

먼저, 분야별 전체 자기소개서의 각각의 단어(명사) 비율은 해당 단어 빈도 수를 전체 단어 빈도수로 나눈 값으로 하여 추출한다. 각 분야에서



〈Figure 3〉 Analysis process

나온 단어의 비율 차이를 바탕으로 각각의 분야에서 내림차순으로 정리하였다. 본 연구는 이 단어들을 각 분야 자기소개서의 차별적 특성의 단어로 정의하였다.

두번째 단계에서는 분야를 구분하지 않은 전체 단어들을 대상으로 LDA를 통해 적절한 토픽 수를 산출하였고, 토픽 수를 바탕으로 첫번째 단계에서 추출한 차별적 특성의 단어들이 어느 토픽에 속하는지를 확인하였다.

세번째 단계에서는 분야별 전체 자기소개서에서 L-LDA를 통해 합격자와 불합격자의 단어와 확률을 추정하였고, 각각의 확률 값의 차이를 바탕으로 이 단어가 합격의 극성에 가까운지, 불합격의 극성에 가까운지를 분류하였다. 세부적인 내용은 3.2.1 ~ 3.2.3절에서 설명하며, 그 과정은 <Figure 3>으로 나타난다.

3.2.1. 지원분야별 차별적 특성의 단어 추출

특정 단어가 전체 문서에서 나타내는 빈도수는 전체 문서에서 특정 단어의 중요성을 파악하는데 용이하다. 이 특정 단어가 다른 집단의 문서집합에서도 중요한지를 비교하기 위해서는 집단 간에 특정 단어의 비율의 차이 값을 산출한다. 따라서 본 연구는 일반분야와 기술분야의 차별적 특성의 단어에 대한 빈도수와 비율의 차이를 산출하였다. 해당분야 전체 단어를 나타내는 집합은 일반분야와 기술분야로 정의한다. 그리고 두 분야 간 각각의 단어별 빈도수를 추출하고 이를 해당 분야의 전체 단어 수로 나눈 것을 word ratio(*wr*)로 정의한다.

$$wr(w,\ D) = \frac{\text{특정 단어의 수 } w}{\text{전체 문서에 나타난 단어의 수 } D}$$

다음으로 특정 단어의 분야별 비율을 비교하기 위해 단어 w에 대한 와 에서의 차이 값을 difference of word ratio(*dwr*)로 정의한다(Kim et al., 2018)

$$dwr(w,\ D_G,\ D_T) = wr(w,\ D_G) - wr(w,\ D_T)$$

Kim et al.(2018)의 *dr*(difference of term ratio)은 특정 단어가 특정 문서에서 나타난 단어들의 비율 $tr(t,\ D)$을 바탕으로 D_a와 D_b 간의 차이로 정의하였으나, 본 연구의 *dwr*(difference of word ratio)은 $wr(w,\ D)$을 바탕으로 D_G와 D_T간의 차이로 정의한다.

마지막으로 일반분야 단어집합(w, dwr(w, D_G, D_T)) 과 기술분야 단어집합(w, dwr(w, D_T, D_G)) 에서 각각의 상위 50개 단어를 통해 두 분야 간에 차별적 특성의 단어를 추출한다.

3.2.2. LDA를 이용한 차별적 특성의 단어에 대한 군집화

3.2.1절에서 추출한 차별적 특성의 단어들이 어떤 토픽에 해당하는 단어인지는 LDA를 통해 나타낼 수 있다. 왜냐하면, 3.1.2절에서 언급한 대로 지원자 간 문항별 실제 자기소개서 토픽이 불일치가 발생하였는데, LDA를 이용하여 자기소개서들이 어떤 문항, 즉, 어떤 토픽에 해당하는지를 분리해 자기소개서를 토픽별로 분리해 줄 수 있기 때문이다.

먼저, 적절한 전체 자기소개서가 몇 개의 토픽으로 나누어지는지를 LDA를 통해 산출한다. Perplexity와 Coherence는 적정한 토픽의 수를 판단하는 지표로, Perplexity 값이 낮고(Blei et al., 2003), Coherence 값이 클수록 실제 문서의 토픽

수를 잘 판단한다고 평가된다(Newman et al., 2010).

두번째, 적절한 토픽 수를 바탕으로 LDA를 이용하여 토픽 및 확률 값을 생성하고, 3.2.1절에서 나타난 차별적 특성의 단어들이 어느 토픽에 속하는지를 추정한다. 차별적 특성의 단어들이 복수의 토픽들에서 중복으로 출현할 수 있는데, 이 때는 단어들의 확률 값이 높은 토픽에 있는 차별적 특성의 단어가 해당 토픽에 속하는 것으로 추정한다.

세번째, 자기소개서 문항의 키워드를 활용하여 라벨링 후보 인덱스로 설정하고, 토픽별 단어 분포들을 고려할 때에 가장 적절한 인덱스를 해당 토픽의 라벨로 설정한다. 또한, 토픽의 라벨로 적절해 보이는 인덱스가 중복으로 나타날 경우, 포괄적 개념을 내포하는 더 상위 범주의 인덱스를 적용하여 토픽의 라벨로 설정한다.

3.2.3. L-LDA를 이용한 차별적 특성의 단어에 대한 합격/불합격 극성 분석

3.2.2절에서 라벨링 된 토픽을 부여한 차별적 특성의 단어에 대해 L-LDA를 이용하여 차별적 특성의 단어가 합격에 가까운지 또는 불합격에 가까운지의 극성을 분석할 수 있다.

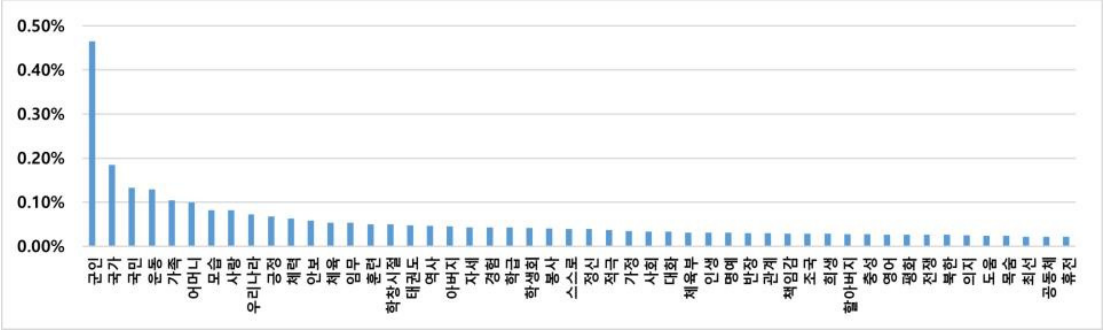
먼저, 자기소개서 및 라벨을 합격/불합격으로 설정한 L-LDA를 이용하여, 각 분야별 합격/불합격 라벨별 토픽 및 확률 값을 생성한다. 그리고 합격/불합격 라벨별로 생성된 토픽 및 확률 값 중 3.2.2절에서 라벨링 된 토픽을 부여한 차별적 특성의 단어만 추출한다.

다음으로 라벨링 된 차별적 특성의 단어별로 합격 라벨에서의 확률 값과 불합격 라벨에서의 확률 값의 차이 값을 추출한다. 이 값이 양의 값이면 합격의 극성을 갖는 것으로 볼 수 있고, 음의 값을 얻으면 불합격의 극성을 갖는 것으로 볼 수 있다.

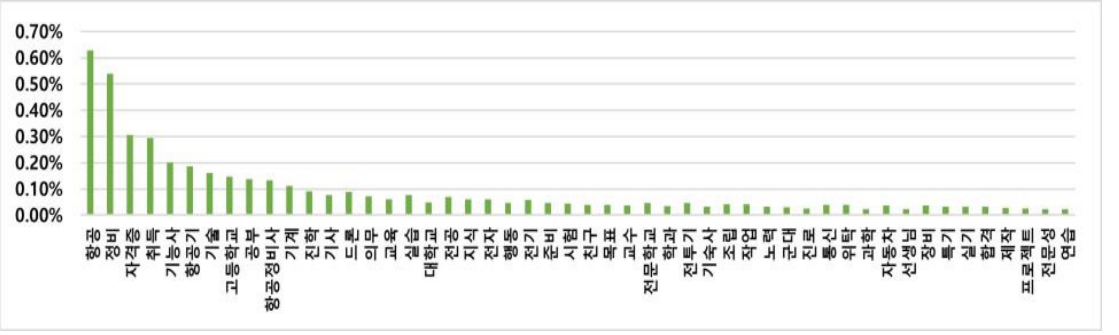
4. 연구결과

4.1. 지원분야별 차별적 특성의 단어 추출

일반분야와 기술분야의 자기소개서 수의 차이로 인해, 단어의 비율을 통해 차별적 특성의 단어 비율을 분석하였다.



〈Figure 4〉 Discriminative characteristic words of general field



〈Figure 5〉 Discriminative characteristic words of technical field

일반분야의 단어로는 주로 가정환경, 성격, 안보관, 학교생활에 대한 단어들이 많이 나왔고, <Figure 4>로 나타난다. 가정환경의 단어로는 “가족”, “어머니”, “아버지”, “가정”, “할아버지”가 나타났다. 성격과 관련된 단어로는 “사랑”, “긍정”, “스스로”, “적극”, “도움”, “관계”, “책임감”, “의지”, “도움”이 나타났다. 안보관과 관련된 단어는 “군인”, “국가”, “국민, “우리나라”, “안보”, “임무”, “훈련”, “역사”, “조국”, “희생”, “충성”, “평화”, “전쟁”, “북한”, “목숨”, “공동체”, “휴전”이 나타났다. 그리고 학교생활과 관련된 단어는 “운동”, “체육”, “학창시절”, “학급”, “학생회”, “봉사”, “체육부”, “반장”으로 나타났다.

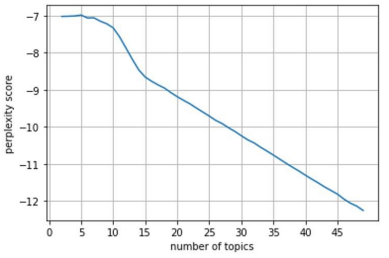
기술분야의 단어로는 주요경력, 학교생활의 단어가 많이 나타났으며 <Figure 5>로 나타난다. 주요경력의 단어로는 “항공”, “정비”, “자격증”, “취득”, “기능사”, “기술”, “항공정비사”, “기계”, “기사”, “전공”, “드론”, “전자”, “시험”, “전문학교”, “전투기”, “통신”, “프로젝트”가 나타났다. 학교생활의 단어로는 “고등학교”, “공부”, “대학교”, “기숙사”, “과학”, “선생님”, “프로젝트”가 나타났다.

차별적 특성의 단어들이 어떤 범주에 들어가는지 분석해보면 일반분야는 단어들이 주로 가정환경, 성격, 안보관, 학교생활에서 나타났고, 기술분야는 주로 주요경력, 학교생활 범주의 단어로 나타났다. 하지만, 이 단어들이 어떤 범주(토픽)에 들어가는지에 대해 좀 더 명확하게 분석해볼 필요가 있고, 일반분야의 “영어”, “체력”이 주요경력의 범주에 들어갈 수 있는지, 학교생활의 범주에 들어갈 수 있는지, 기술분야의 “시험”, “실기”라는 단어가 주요경력이라는 범주인지 학교생활의 범주인지에 대해서 4.2절에서 분석한다.

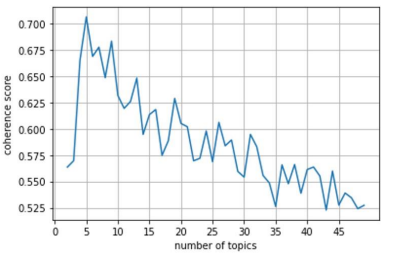
4.2. LDA를 이용한 차별적 특성의 단어에 대한 군집화

4.1절에서 나타난 차별적 특성의 단어들을 바탕으로 자기소개서를 범주화 할 수 있는 수, 범주안에 들어가는 단어를 LDA로 분석할 수 있다.

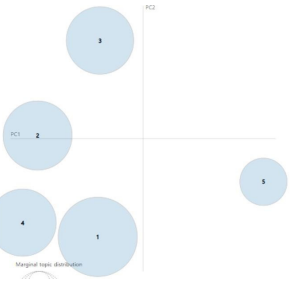
먼저, LDA의 적절한 토픽 수를 확인하기 위해 Perplexity와 Coherence 점수를 평가하였다. 분석결과, 5개의 토픽 수에서 Perplexity 점수는 -7.01점, Coherence 점수는 0.71점으로 가장 좋은 점수를 나타냈고, <Figure 6>, <Figure 7>, <Figure 8>로 나타난다.



〈Figure 6〉 Perplexity score



〈Figure 7〉 Coherence score



〈Figure 8〉 Appropriate number of Topics

〈Table 2〉 The result of LDA topic modeling for cover letter

Topic 1 (28.1%)		Topic 2 (21.3%)		Topic 3 (20.4%)		Topic 4 (20.2%)		Topic 5 (10%)	
활동	2.87%	국가	3.82%	부사관	5.50%	부모님	6.28%	친구	2.40%
고등학교	2.41%	국민	2.51%	공군	4.42%	아버지	5.86%	노력	1.88%
동아리	2.32%	좌우명	1.94%	지원	2.03%	어머니	3.58%	성격	1.81%
학교	2.27%	북한	1.66%	항공	1.93%	가족	3.23%	시간	1.38%
경험	1.54%	우리나라	1.63%	직업	1.92%	가정	2.05%	다른	1.31%
봉사	1.24%	대한민국	1.62%	자격증	1.62%	말씀	1.77%	먼저	1.25%
중학교	1.24%	군인	1.34%	취득	1.35%	사랑	1.09%	자신	1.21%
선생님	1.23%	안보	1.29%	정비	1.31%	동생	1.04%	행동	1.19%
봉사활동	1.09%	노력	1.28%	관심	1.00%	모습	0.95%	장점	1.08%
학생	1.05%	전쟁	1.15%	노력	0.80%	환경	0.86%	배려	1.07%
:	:	:	:	:	:	:	:	:	:

〈Table 3〉 The Questions in the cover letter

No.	Question	Detailed Questions
Q 1	**Home environment** - *Topic 4*, **Growth environment**	The advantages and disadvantages of family
Q 2	**Growth process** (School life - *Topic 1*, Club activities, Volunteer work, Student Council Experience)	The most rewarding experience, The most difficult experience
Q 3	**Self-expression** (Personality - *Topic 5*, View of state, Motto, Values, View of national security - *Topic 2)*	Strengths and weaknesses
Q 4	**Reason for application, Vision, Aspiration**	Major career - *Topic 3*, Research performance, Cerficicate, Foreign language ability

두번째, 5개 토픽의 LDA를 분석한 결과에 따른 단어들은 다음의 <Table 2>와 같다. <Table 2>는 LDA에서 나온 결과들의 단어들 중 차별적 특성의 단어들에 대해 순서대로 정렬하였다. 만

〈Table 4〉 Clustering of discriminative characteristic words using LDA for general field

Home environment		School life		체육부	0.14%	View of national security		평화	0.47%
아버지	5.86%	경험	1.54%	Personality		국가	3.82%	휴전	0.41%
어머니	3.58%	봉사	1.24%	도움	0.80%	북한	1.66%	조국	0.37%
가족	3.23%	운동	0.81%	긍정	0.63%	우리나라	1.63%	목숨	0.36%
가정	2.05%	학급	0.68%	스스로	0.61%	군인	1.34%	임무	0.29%
사랑	1.09%	반장	0.63%	관계	0.55%	안보	1.29%	훈련	0.24%
모습	0.95%	학생회	0.44%	자세	0.41%	전쟁	1.15%	충성	0.12%
책임감	0.77%	부장	0.33%	사회	0.32%	희생	0.67%	명예	0.09%
할아버지	0.48%	체육	0.26%	적극	0.28%	최선	0.61%	Major career	
인생	0.30%	학창시절	0.23%	공동체	0.19%	역사	0.59%	체력	0.42%
대화	0.24%	태권도	0.19%	의지	0.14%	정신	0.49%	영어	0.26%

〈Table 5〉 Clustering of discriminative characteristic words using LDA for technical field

Home environment		기숙사	0.15%	View of national security		시험	0.57%	전문성	0.29%
교육	0.43%	제작	0.15%	군대	0.33%	기능사	0.52%	기사	0.27%
진로	0.15%	교수	0.12%	의무	0.18%	진학	0.50%	전투기	0.26%
School life		조립	0.11%	Major career		항공기	0.45%	전기	0.16%
고등학교	2.41%	프로젝트	0.10%	항공	1.93%	합격	0.45%	전자	0.15%
공부	0.74%	Personality		자격증	1.62%	드론	0.44%	장비	0.15%
연습	0.56%	친구	2.40%	취득	1.35%	지식	0.40%	전문학교	0.14%
대학교	0.44%	노력	1.88%	정비	1.31%	전공	0.35%	실기	0.12%
학과	0.26%	행동	1.19%	목표	0.68%	항공정비사	0.32%	통신	0.11%
실습	0.23%	약속	0.29%	준비	0.67%	기계	0.31%	자동차	0.09%
과학	0.21%	선생님	0.15%	기술	0.61%	특기	0.30%	작업	0.09%

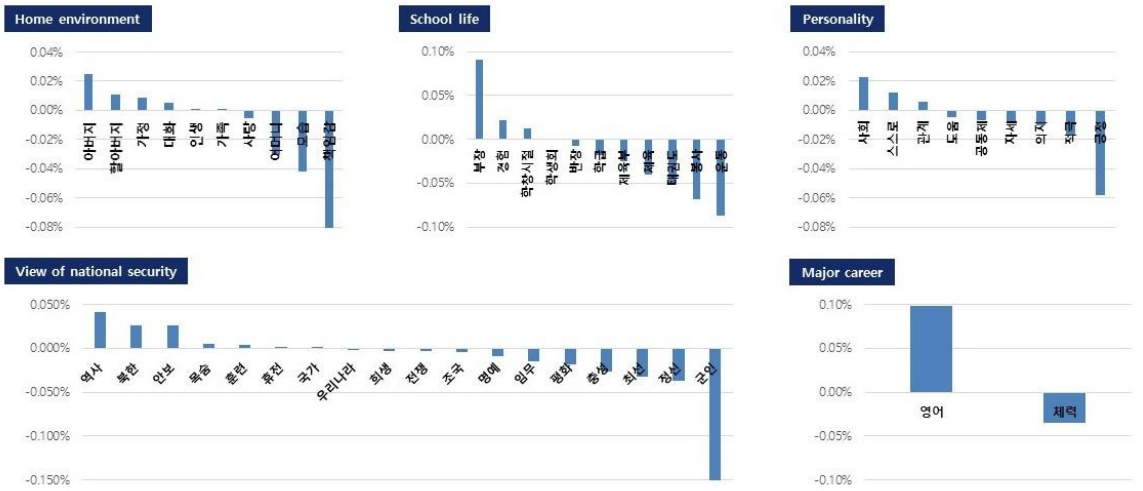
약, 차별적 특성의 단어들이 복수의 토픽에서 나타났다면, 확률 값이 높은 토픽에서 나온 단어를 그 토픽으로 설정하였다.

세번째, 주어진 단어들을 바탕으로 토픽을 라벨링하였다. 라벨링에 쓰인 단어는 <Table 3>에 나타난 자기소개서에서 문항의 키워드 중 토픽 내 단어들을 대표할 수 있는 가장 유사한 키워드를 선정하였다.

군집화 된 단어들은 <Table 4>에서 일반분야, <Table 5>에서 기술분야로 나타난다.

4.3. L-LDA를 이용한 차별적 특성의 단어에 대한 합격/불합격 극성 분석

4.2절에서 LDA를 통해 차별적 특성의 단어들에 대해 범주화하였다면, L-LDA는 합격, 불합격의 라벨을 통해 극성을 파악할 수 있다. 그리고



〈Figure 9〉 Polarities of discriminative characteristic words using L-LDA for the general field

L-LDA로 발현된 단어의 확률 값의 차이를 통해 각 단어들이 합격에 가까운 단어인지, 불합격에 가까운 단어인지를 확인할 수 있다. 확률 값의 차이가 양의 값이 나온 경우는 합격자 자기소개서의 성향에 더 가깝다고 할 수 있고, 확률 값의 차이가 음의 값이 나온 경우는 불합격자 자기소개서의 성향에 더 가깝다고 할 수 있다.

먼저, 일반분야의 결과는 <Figure 9>와 같으며, 가정환경(Home environment)에서는 “아버지”, “할아버지”, “가정”, “대화” 등에 합격자 자기소개서의 성향이 나타났고, “책임감”, “모습”, “어머니” 등에 불합격자 자기소개서의 성향이 나타났다. 학교생활(School life)에서는 “부장”, “경험”, “학창시절” 등에 합격자 자기소개서의 성향이 나타났고, “운동”, “봉사”, “태권도”, “체육부”, “학급”, “반장” 등에 불합격자 자기소개서의 성향이 나타났다. 성격(Personality)에서는 “사회”, “스스로”, “관계”에 합격자 자기소개서의 성향이 나타났고, “긍정”, “적극”, “의지”, “자세”, “공동체” 등에 불합격자 자기소개서의 성향이 나타났다. 안보관(View of national security)에서는 “역사”, “북한”, “안보” 등에 합격자 자기소개서의 성향이 나타났고, “군인”, “정신”, “최선”, “충성”, “평화”, “임무” 등에 불합격자 자기소개서의 성향이 나타났다. 주요경력(Major career)에서는 “영어”에 합격자 자기소개서의 성향이 나타났고, “체력”에 불합격자 자기소개서의 성향이 나타났다.

다음으로, 기술분야의 결과는 <Figure 10>과 같으며, 가정환경(Home environment)에서는 “진로”, “교육”에 불합격자 자기소개서의 성향이 나타났다. 학교생활(School life)에서는 “공부”, “기숙사”, “연습”, “과학” 등에 합격자 자기소개서의 성향이 나타났고, “고등학교”, “조립”, “제작”, “실습”, “대학교”에 불합격자 자기소개서의 성향이 나타났다. 성격(Personality)에서는 “노력”에 합격자 자기소개서의 성향이 나타났고, “친구”, “선생님”, “약속”에 불합격자 자기소개서의 성향이 나타났다. 그리고 안보관(View of national security)에서는 “군대”에 합격자 자기소개서의

〈Figure 10〉 Polarities of discriminative characteristic words using L-LDA for the technical field

성향이 나타났고, “의무”에 불합격자 자기소개서의 성향이 나타났다. 주요경력(Major career)에서는 “항공”, “취득”, “항공정비사”, “목표”, “정비”, “기사” 등에 합격자 자기소개서의 성향이 나타났고, “드론”, “자동차” 등에 불합격자 자기소개서의 성향이 나타났다.

5. 결론

본 연구는 단어 비율의 차이, LDA, L-LDA를 이용하여 공군 부사관 자기소개서에서 차별적 특성의 단어들에 대해 토픽과 합격 및 불합격의 극성을 추정하였다. 이는 복잡한 수학적 모델 또는 신경망 모델을 사용하지 않고도, 자기소개서

〈Table 6〉 Contexts of word in the cover letter of successful applicant

Field	Word	Context
General field	English (영어)	도내 영어(English) 말하기 대회에 참가하여 많은 학생들과 선의의 경쟁을 할 수 있는 기회를 얻었으며, 영어듣기평가 대회, 뮤지컬, 팝송대회 등 동아리 친구들과 함께 협동 정신을 기를 수 있는 경험도 얻을 수 있었습니다.
	Leader (부장)	학교에서는 독서토론동아리, 역사동아리와 같이 다양한 동아리에서 부장(Leader)을 맡아 리더십을 키워왔습니다.
Technical field	Aviation (항공)	항공(Aviation)에 관심이 생기며 공군에 대한 관심이 높아지게 되었고 전투복을 입으며 생활하고 싶은 꿈이 생기게 되었습니다.
	Certification (취득)	공군 부사관이 되기 위하여 매일 3 시간씩 공부하며 항공기체정비기능사, 항공기관정비기능사, 에너지관리기능사 등을 취득(Certification)했습니다.
	Flight mechanic (항공정비사)	공군 부사관으로서 국방의 의무를 충실히 잘 해내고 항공기 정비를 통해 항공정비사(Flight mechanic)에 필요한 자질, 기술 그리고 많은 경험과 경력 등을 쌓아 더 나은 미래를 위해 발전하며 나아가는 목표를 가지고 지원하였습니다.

〈Table 7〉 Contexts of word in the cover letter of unsuccessful applicant

Field	Word	Context
General field	Serviceperson (군인)	대한민국에 태어난 것을 자랑스럽게 여기며 자랑스러운 대한민국 공군이 되어 나라에 기여할 수 있는 훌륭한 군인(Serviceperson)이 되겠습니다.
	Responsibility (책임감)	부사관으로 임관하여 병사와 같이 통제를 받는 게 아닌 간부로서 통제를 하는 사람이 되어 책임감(Responsibility) 있는 일을 할 것입니다.
	Sports (운동)	초등학교 때부터 고등학교때까지 운동(Sports)을 좋아하여 운동부에 들어가서 운동을 하였습니다.
Technical field	Drone (드론)	저는 저의 드론(Drone) 비행 실력으로 저의 국가의 국민이 외부세력의 위협으로 부터 안전한 삶을 보장받을 수 있도록 만들고 싶어 지원하게 되었습니다.
	Obligation (의무)	남자라면 가야하는 군대를 공군 부사관으로 국방의 의무(Obligation)를 이행하여 나라를 지키고 싶습니다.

의 합격 및 불합격 성향을 추정할 수 있었다.

일반분야는 안보관에서 기술분야는 주요경력에서 가장 많은 차별적 특성의 단어가 나타남을 알 수 있다. 또한, 합격에 가까운 단어는 <Table 6>와 같으며, 일반분야에서는 “영어”, “부장”이 나타났고 기술분야에서는 “항공”, “취득”, “항공정비사” 등이 나타났다. 반대로 불합격에 가까운 단어는 <Table 7>과 같으며, 일반분야에서는 “군인”, “책임감”, “운동”이 나타났고, 기술분야에서는 “드론”, “의무” 등이 나타났다.

본 연구는 일반분야와 기술분야의 자기소개서에서 각각 50개의 차별적 특성의 단어를 추출하였다. 일반분야에서는 50개 단어 중 21개 단어가 합격에 가까운 단어로 분리되었고, 기계분야에서는 50개 단어 중 35개 단어가 합격에 가까운 단어로 분리되었다. 합격의 Independent Sample T-Test 결과, 일반분야 합격자 평균은 15.1%, 기술분야 합격자 평균은 23.1%이며, 95% 신뢰수준에서 t = -15.935, p = 7.33E-57으로 p가 0.001보다 작아 귀무가설인 두집단의 평균은 같다가 기각되어 두 분야는 차이가 있는 것으로 나타난다. 또한, 이 결과는 기술분야의 더 높은 합격자 평균이 기술분야 자기소개서에서 더 많은 합격 극성의 단어들을 추정한 본 연구의 결과와 일치한다. 따라서 본 연구의 분석절차는 합격자와 불합격자의 차별적 특성을 식별하는 것에 유용할 것으로 판단된다.

5.1. 실무적 시사점

첫번째, 민간분야의 자기소개서와 다른 공군 부사관 지원자의 자기소개서에 대한 특성을 반영한 최초의 연구이다. 기업 또는 대학교의 지원을 위한 자기소개서와 달리 직업군인 또는 병역의 의무를 위해 지원하는 점에서 다른 특성을 가진다. 또한, 4년제 대학교 졸업자에 한해 지원할 수 있는 장교와 달리 부사관은 고등학교 졸업자, 대학 이상 재학 또는 졸업자가 지원하기 때문에, 경력과 학위에 있어서도 다양한 형태를 나타낸다. 이에 대한 연구를 진행했다는 점에 의의가 있다.

두번째, 연구에서 제안한 방법론이 군 인사분야에 있는 다른 형태의 대량 문서에도 적용 가능하다. 군 인사분야 에는 차별적 특성과 극성을 분리하는 분석이 필요한 분야가 다수 있다. 진급추천 심사에 제출되는 복무계획서, 지휘관 의견

서, 근무평정에 제출되는 업무실적, 군사전문성, 조직기여/활성화 부분, 사건사고 예방을 위한 면담기록부 등이 있다. 진급 선발자와 비선발자, 상위 근무평정자와 하위 근무평정자, 사고자와 비사고자를 대상으로 어떤 차별적 특성이 있는지를 추출할 수 있다. 이를 통해, 군 인사분야에 있어 심층적인 분석을 가능하게 해준다.

세번째, 우수한 군 간부획득을 위한 군 인력획득분야에 활용 가능한 정보를 제공한다. 예를 들면, 일반분야의 차별적 속성의 단어는 주요 경력이 50개 중 2개이고, 기술분야의 차별적 속성의 단어는 가정환경과 안보관이 50개 중 각 2개라는 점은 지원분야별로 더 면밀한 관찰이 필요하다는 정보를 제공한다. 그리고 어학능력, 리더십, 자격증 등의 단어가 합격 극성이 두드러진 점은 직무능력에 대한 선발평가 잘 이루어지고 있음과 동시에, 직무능력 외의 평가요소 관련 보완필요성에 대한 정보를 제공한다. 또한, 공군 부사관의 53개 특기 중 특정 특기들에 대한 단어들이 다수 나타남에 따라, 다양한 재능을 보유한 자원들이 공군 부사관에 지원할 수 있는 홍보정책의 필요성에 대한 정보를 제공한다. 이 외에도 많은 정보를 실무에 활용할 수 있다.

5.2. 이론적 시사점

본 연구를 통해, 자기소개서의 차별적 특성을 추출하기에 L-LDA 모델이 더 적합하다는 것이 잘 나타난다. 비지도학습인 LDA 대량의 문서를 적절한 토픽 수를 산출해주고, 토픽 수를 바탕으로 적절하게 문서를 분류해줄 수 있다. 그러나 L-LDA는 라벨이 있는 문서들에서 그 라벨을 추정하는 토픽이 되는 단어들을 추출해준다. 따라서 합격, 불합격의 라벨을 가진 문서들을 분리하기에는 L-LDA가 더 적절하다.

또한, LDA와 L-LDA를 조합하여 사용하는 방법을 제시한다. LDA는 토픽을 문항의 분류로 활용할 수 있고, L-LDA는 합격, 불합격 분류로 활용할 수 있다. 이처럼 대량의 문서에서 토픽별 분류와 라벨별 분류가 필요하다면, LDA와 L-LDA를 각각 적용해 볼 수 있다.

5.3. 한계점 및 향후 연구방향

본 연구에 사용된 데이터는 실제 데이터로서 개인정보 비식별화 조치를 실시하였다. 그러나 지원접수 시에 자기소개서 제출은 현재도 지속적으로 시행되고 있는 민감한 부분이다. 그리고 본 연구는 2019년도 공군 부사관후보생 지원자들의 자기소개서를 분석하였기 때문에, 연구의 결과를 전체 공군 부사관후보생 지원자의 성향으로 일반화하는 것은 지원자들에게 혼란을 초래할 수 있다. 따라서 4장의 연구결과를 향후에 다른 방법론과 주제로 해석 및 연구하고자 한다. 하지만, 인사선발 담당자는 본 연구결과에서 다수의 의미 있는 정보들을 해석할 수 있다.

또한, 공군 부사관후보생 지원자 자기소개서는 합격을 좌우하는 직접적인 평가요소가 아닌 2차 전형의 면접을 위한 참고자료라는 한계점이 있다. 하지만, 2차 전형의 면접을 위해 지원자들은 정제된 언어를 사용하여 자신의 성향이 드러나는 자기소개서를 작성한다는 점과 공군 부사관후보생 합격자와 불합격자 자기소개서의 차별적 특성을 추출한 최초의 연구결과라는 점을 바탕으로, 본 연구결과는 공군 부사관후보생 선발정책과 홍보정책에서 유의미한 분석자료로 충분히 활용할 수 있고, 연도 단위로 선발결과를 정리하면 공군 부사관후보생 합격자 및 불합격자

성향의 변화 추이도 관찰할 수 있다.

그리고 연구결과는 여군 지원자에 대한 합격/불합격 성향을 충분히 반영하지 못한 편견(Bias) 문제가 있다. 예를 들면, 여군 지원자들의 자기소개서에서는 가정환경과 관련된 "오빠", "언니"와 같은 단어도 나타났는데, '20년 기준 여군 간부 비중이 7.5%임을 고려할 때에 여군 지원자들의 특성이 반영되는 단어들이 충분히 반영되지 않았을 가능성이 있다. 향후, 연구에서는 여군 지원자의 합격/불합격 성향도 충분히 반영되었는지에 대한 검토가 필요하다.

또한, 지원자의 임관 후 데이터 사용을 하지 못한 실무적인 한계가 있었다. 향후, 연구과제로서 임관 후에 인사데이터까지 적용할 수 있다면, 지원자들의 장기복무 선발여부, 진급여부, 징계여부 등의 면밀한 분석 및 예측모델 구축이 가능하다. 이를 통해 우수한 군 간부선발에 관한 정책 검토 및 제안에 더욱 기여할 수 있다.

참고문헌(References)

Allahyari, M., S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, K. Kochut, “A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques,” *arXiv: 1707.02919v2*(2015).

Bae, S. H., X. Ku, C. Park, J. Ki, “A Latent Topic Modeling approach for Subject Summarization of Research on the Military Art and Science in South Korea,” *Korean Journal of Military Art and Science*, Vol.76, No.2(2020), 181 ~ 216.

Baek, S. Y., J. U. Leem, H. J. Kwon, “An Empirical Study on The Relationship Between Professional Soldiers Selection Variables and Job Satisfaction, Job Performance,” *Journal of Employment and Career*, Vol.9, No.2(2019), 95-116.

Blei, D. M., A. Y. Ng, M. I. Jordan, “Latent dirichlet allocation,” *Journal of Machine Learning Research*, Vol.3(2003), 993-1022.

Dohkgoh, S., P. R. Kim, “The deepening of low birthrates and the issue of military manpower acquisition in developed countries,” *KIDA Defense Weekly*, Vol.1652(2017).

Jeon, G. W., I. Kang, J. H. Jeon, "Systematic Analysis on the Trend of Defense Technologies Using Topic Modeling : A Case of an Armoured Fighting Vehicle,” *The Journal of Business and Economics*, Vol.36, NO.1(2020), 69-94.

Kim, D. W., J. Y. Kang, J. I. Lim, “Comparative Analysis of Job Satisfaction Factors, Using LDA Topic Modeling by Industries : The Case Study of Job Planet Reviews,” *Journal of Information Technology Services*, Vol.15, No.3(2016), 157-171.

Kim, H. J., W. J. Kim, "A Study on Automatic Analysis System of National Defense Articles,” *Journal of the KIMST*, Vol.21, No.1(2018), 86-93.

Kim, H. K., “A Study on Teaching How to Write a Cover Letter for a Job,” *The Society Of Korean Literary Criticism*, Vol.51(2014), 7-34.

Kim, S. G., J. Y. Kang, “Analyzing the discriminative attributes of products using text mining focused on cosmetic reviews,” *Information Processing and Management*, Vol.54, No.6(2018), 938-957.

Kim, Y. S., H. S. Moon, J. K. Kim, “Self

Introduction Essay Classification Using Doc2Vec for Efficient Job Matching," *Journal of Information Technology Service*, Vol.19, No.1(2020), 103-113.

Lee, C. Y., H. S. Moon, "Study on analysis of North Korea`s news trends associated with provocations using text mining," *Journal of National Defence Studies*, Vol.59, No.4(2016), 103-124.

Lee, D. G., I. H. Kim, "An Analysis of Self-introduction Texts based on Statistical Text Analysis," *Korean Cultural Studies*, Vol.81(2018), 649-684.

Lee, J. H., S. H. Jung, J. H. Kim, E. J. Min, U. Y. Yeo, J. W. Kim, "Product Evaluation Criteria Extraction through Online Review Analysis : Using LDA and k-Nearest Neighbor Approach," *Journal of Intelligence and Information Systems*, Vol.26, No.1(2020), 97-117.

Lee, M. C., H. J. Kim, "Construction of Event Networks from Large News Data Using Text Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.24, No.1(2018), 183-203.

Lim, S. S., M. G. Lee, "A study on military organizational tasks analysis methodology," *The Korean Data and Information Science Society*, Vol.30, No.1(2019), 139-157.

Moon, S. H., J. Y. Kang, "A study on detective story authors' style differentiation and style structure based on Text Mining," *Journal of Intelligence and Information Systems*, Vol.25, No.3(2019), 89-115.

Newman, D., J. H. Lau, K. Grieser, T. Baldwin, "Automatic evaluation of topic coherence," *In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* (2010), 100-108.

Oh, S. H., H. J. Kim, "A Study on the 'Low Fertility' Research Trends Using Text Mining Technique: Focusing on the Comparison with the Process of Low Fertility Policy," *Health and Social Welfare Review*, Vol.40, No.3 (2020), 492-533.

Ramage, D., D. Hall, R. Nallapati and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," *Proceedings of the 2009 conference on empirical methods in natural language processing,* (2009).

Shin, J. S., "A Study on Teaching Method of Self-introduction for Employment," *A collection of Southeast Asian literature*, Vol.40(2015), 83-113.

Tan, A. H., "Text mining: The state of the art and the challenges," *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*, (1999), 65-70.

Teh, Y. W., M. I. Jordan, M. J. Beal and D. M. Blei, "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, Vol.101, No.476(2006), 1566-1581.

Yoon, S., S. Kim, K. Shin, "Development of the Accident Prediction Model for Enlisted Men through an Integrated Approach to Datamining and Textmining," *Journal of Intelligence and Information Systems*, Vol.21, No.3(2015), 1-17.

Abstract

Analyzing the discriminative characteristic of cover letters using text mining focused on Air Force applicants

Hyeok Kwon* · Wooju Kim**

The low birth rate and shortened military service period are causing concerns about selecting excellent military officers. The Republic of Korea entered a low birth rate society in 1984 and an aged society in 2018 respectively, and is expected to be in a super-aged society in 2025. In addition, the troop-oriented military is changed as a state-of-the-art weapons-oriented military, and the reduction of the military service period was implemented in 2018 to ease the burden of military service for young people and play a role in the society early. Some observe that the application rate for military officers is falling due to a decrease of manpower resources and a preference for shortened mandatory military service over military officers. This requires further consideration of the policy of securing excellent military officers. Most of the related studies have used social scientists' methodologies, but this study applies the methodology of text mining suitable for large-scale documents analysis. This study extracts words of discriminative characteristics from the Republic of Korea Air Force Non-Commissioned Officer Applicant cover letters and analyzes the polarity of pass and fail. It consists of three steps in total. First, the application is divided into general and technical fields, and the words characterized in the cover letter are ordered according to the difference in the frequency ratio of each field. The greater the difference in the proportion of each application field, the field character is defined as 'more discriminative'. Based on this, we extract the top 50 words representing discriminative characteristics in general fields and the top 50 words representing discriminative characteristics in technology fields. Second, the number of appropriate topics in the overall cover letter is calculated through the LDA. It uses perplexity score and coherence score. Based on the appropriate number of topics, we then use LDA to generate topic and probability, and estimate which topic words of discriminative characteristic belong to. Subsequently, the keyword indicators of questions used to set the labeling candidate index, and the most appropriate index indicator is set as

* Department of Industrial Engineering, Yonsei University
** Corresponding Author: Wooju Kim
Department of Industrial Engineering, Yonsei University
50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: +82-2-2123-7754, E-mail: wkim@yonsei.ac.kr

the label for the topic when considering the topic-specific word distribution. Third, using L-LDA, which sets the cover letter and label as pass and fail, we generate topics and probabilities for each field of pass and fail labels. Furthermore, we extract only words of discriminative characteristics that give labeled topics among generated topics and probabilities by pass and fail labels. Next, we extract the difference between the probability on the pass label and the probability on the fail label by word of the labeled discriminative characteristic. A positive figure can be seen as having the polarity of pass, and a negative figure can be seen as having the polarity of fail. This study is the first research to reflect the characteristics of cover letters of Republic of Korea Air Force non-commissioned officer applicants, not in the private sector. Moreover, these methodologies can apply text mining techniques for multiple documents, rather survey or interview methods, to reduce analysis time and increase reliability for the entire population. For this reason, the methodology proposed in the study is also applicable to other forms of multiple documents in the field of military personnel. This study shows that L-LDA is more suitable than LDA to extract discriminative characteristics of Republic of Korea Air Force Noncommissioned cover letters. Furthermore, this study proposes a methodology that uses a combination of LDA and L-LDA. Therefore, through the analysis of the results of the acquisition of non-commissioned Republic of Korea Air Force officers, we would like to provide information available for acquisition and promotional policies and propose a methodology available for research in the field of military manpower acquisition.

Key Words : Air Force Non-Commissioned Officer, Cover Letter, Text Mining, LDA, L-LDA

Received : May 21, 2021 Revised : July 23, 2021 Accepted : July 29, 2021
Corresponding Author : Wooju Kim

저 자 소 개

권 혁
대한민국 공군 소령이며, 연세대학교 정경대학 경영학과에서 경영학사와 문과대학 노어노문학과에서 문학사를 취득하였다. 현재 군 위탁생으로서 연세대학교 일반대학원 산업공학과에서 석사과정으로 재학 중이다. 관심분야는 자연어 처리, 머신 러닝 등이다.

김 우 주
1987년 연세대학교 BBA 과정 학사 학위를 취득하였고, 1994년 KAIST 경영과학 박사를 취득하였으며, 현재 연세대학교 산업공학과 교수로 재직 중이다. 관심분야는 시맨틱 웹, 시맨틱 웹 환경의 의사결정지원 시스템, 시맨틱 웹 마이닝, 지식관리 및 인공지능 웹 서비스이다.