자연어 처리 및 기계학습을 활용한 제조업 현장의 품질 불량 예측 방법론

[1]노정민, [2*]김용성

A Method for Prediction of Quality Defects in Manufacturing Using Natural Language Processing and Machine Learning

[1]Jeong-Min Roh, [2*]Yongsung Kim

요 약

제조업 현장에서 제작 공정 수행 전 품질 불량 위험 공정을 예측하여 사전품질관리를 수행하는 것은 매우 중요한 일이다. 하지만 기존 엔지니어의 역량에 의존하는 방법은 그 제작공정의 종류와 수가 다양할수록 인적, 물리적 한계에 부딪힌다. 특히 원자력 주요기기 제작과 같이 제작공정이 매우 광범위한 도메인 영역에서는 그 한계가 더욱 명확하다. 본 논문은 제조업 현장에서 자연어 처리 및 기계학습을 활용하여 품질 불량 위험 공정을 예측하는 방법을 제시하였다. 이를 위해 실제 원자력발전소에 설치되는 주기기를 제작하는 공장에서 6년 동안 수집된 제작 기록의 텍스트 데이터를 활용하였다. 텍스트 데이터의 전처리 단계에서는 도메인 지식이 잘 반영될 수 있도록 단어 사전에 Mapping 하는 방식을 적용하였고, 문장 벡터화 과정에서는 N-gram, TF-IDF, SVD를 결합한 하이브리드 알고리즘을 구성하였다. 다음으로 품질 불량 위험 공정을 분류해내는 실험에서는 k-fold 교차 검증을 적용하고 Unigram에서 누적 Trigram까지 여러 케이스로 나누어 데이터셋에 대한 객관성을 확보하였다. 또한, 분류 알고리즘으로 나이브 베이즈(NB)와 서포트 벡터 머신(SVM)을 사용하여 유의미한 결과를 확보하였다. 실험결과 최대 accuracy와 F1-score가 각각 0.7685와 0.8641로서 상당히 유효한 수준으로 나타났다. 또한, 수행해본 적이 없는 새로운 공정을 예측하여 현장 엔지니어들의 투표와의 비교를 통해서 실제 현장에 자연스럽게 적용할 수 있음을 보여주었다.

Abstract

Quality control is critical at manufacturing sites and is key to predicting the risk of quality defect before manufacturing. However, the reliability of manual quality control methods is affected by human and physical limitations because manufacturing processes vary across industries. These limitations become particularly obvious in domain areas with numerous manufacturing processes, such as the manufacture of major nuclear equipment. This study proposed a novel method for predicting the risk of quality defects by using natural language processing and machine learning. In this study, production data collected over 6 years at a factory that manufactures main equipment that is installed in nuclear power plants were used. In the preprocessing stage of text data, a mapping method was applied to the word dictionary so that domain knowledge could be appropriately reflected, and a hybrid algorithm, which combined n-gram, Term Frequency–Inverse Document Frequency, and Singular Value Decomposition, was constructed for sentence vectorization. Next, in the experiment to classify the risky processes resulting in poor quality, k-fold cross-validation was applied to categorize cases from Unigram to cumulative Trigram. Furthermore, for achieving objective experimental results, Naive Bayes and Support Vector Machine were used as classification algorithms and the maximum accuracy and F1-score of 0.7685 and 0.8641, respectively, were achieved. Thus, the proposed method is effective. The performance of the proposed method were compared and with votes of field engineers, and the results revealed that the proposed method outperformed field engineers. Thus, the method can be implemented for quality control at manufacturing sites.

Keywords: *Manufacturing, Natural Language Processing, Machine Learning, Prediction of Quality Defects, Sentence Vectorization*

[1]*고려사이버대학교 융합정보대학원(jmnoh1027@cuk.edu)*
[2 교신저자]*고려사이버대학교 창의공학부 소프트웨어공학과 조교수(kys1001@cuk.edu)*
Received: Sept. 02, 2021, Revised: Sept. 21, 2021, Accepted: Sept. 21, 2021

I. 서론

제조업 현장의 품질 관리는 원가, 생산성 그리고 납기에 영향을 미치기 때문에 매우 중요하다. 특히 원자력 주요기기 제작과 같이 높은 수준의 품질을 요구하면서 하나의 프로젝트가 수년에 걸쳐 진행되는 장기수주장치산업 분야에서는 그 중요성이 더욱 크다. 이곳에서는 모든 제작 과정이 규제기관에서 정한 Code 에 따라 수행되므로 아무리 사소해 보이는 품질 문제라도 일단 발생하면 그 즉시 제작을 중단하고 문제를 해결한 다음에 제작을 진행할 수 있기 때문이다. 그리고 품질 관리를 효과적으로 수행하기 위해서는 무엇보다 사전에 품질 불량 발생을 예측하는 것이 중요하다. 제작공정을 수행하기 전에 미리 품질 불량 위험 공정을 파악하여 선제적으로 대응할 수 있기 때문이다. 하지만 엔지니어의 도메인 역량에 의존하는 기존의 전통적인 방식은 한계가 있다. 모든 제작공정을 일정한 기준에 맞춰 검토하는 것이 현실적으로 가능하지 않기 때문이다. 더군다나 원자력 주요기기 제작과 같이 제작공정이 매우 다양하고 그 수가 많은 대형 제조업 프로젝트에서는 엔지니어가 도메인 역량을 쌓기까지 짧지 않은 시간이 걸리므로 그 한계는 더욱 분명하다.

본 논문은 이러한 한계를 극복하기 위해 자연어 처리와 기계학습을 활용하여 품질 불량 발생 위험 공정을 사전에 예측하는 방법을 제안한다. 그림 1 은 그 원리이며 하나의 인과관계를 나타낸다. 공정을 설명하는 텍스트는 실제 작업을 표현하고, 실제 작업은 품질 불량의 유형 및 가능성과 관계가 있다. 따라서 공정 설명 텍스트를 과거의 품질 불량 발생 여부에 따라 라벨링을 한 다음 임베딩하여 벡터공간에 사상하면, 기계학습의 분류 방식으로 품질 불량이 발생할지를 예측하는 것이 가능하다.

본 논문의 구성은 다음과 같다. II 장에서는 관련 연구를 소개하고, III 장에서는 본 논문에서 제안하는 방법에 관하여 설명한다. 그리고 IV 장에서는 제안한 기법에 대한 실험결과를 정리하고, V 장에서는 결론을 맺는다.

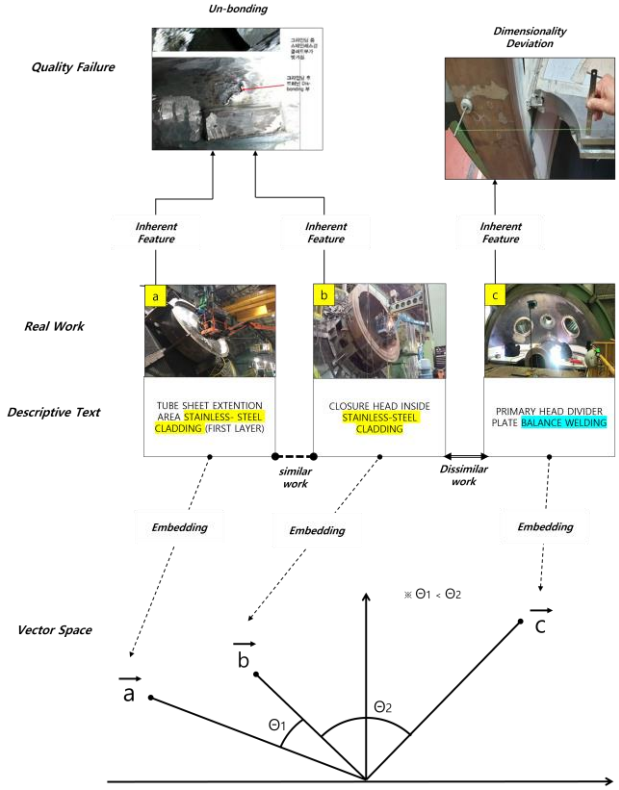

Figure 1. Relations among defect, real work, distribution text and embedding vector
그림 1. 품질 불량, 실제 작업, 작업 설명 텍스트 그리고 임베딩 벡터 간의 관계

II. 관련 연구

이 장에서는 제조업 현장에서 자연어 처리 또는 기계학습을 활용한 연구들과 자연어 처리 및 기계학습과 관련된 연구들을 살펴본다.

2.1 제조업 현장에서 자연어 처리 또는 기계학습을 활용한 연구

제조 산업군에 종사하는 기업들은 현장에서 발생하는 공정 데이터를 이용하여 품질 불량에 대한 의사 결정에 활용하고 있다. 여기에서 객관적인 의사 결정의 지표가 필요한데 스마트 공장에서 생성되는 대규모의 데이터들을 분석하기 위해서는 선형 회귀분석과 같은 예전의 분석방법 외에 다양한 기계학습 방법들이 필요하다[1]

설비 고장에서 패턴을 발견하는 것 또한 제조업에서 자연어 처리 및 기계학습을 활용하는 목적 중 하나이다. [2]는 설비 오류 텍스트 데이터를 분석하고 연관 규칙 마이닝으로 설비에서 오류가 발생하는 패턴을 도출해내는 프레임워크를 제안하였다. 구체적으로, 설비 오류 이력은 한 단어만으로는 전문 용어의 의미를 표현하기 어렵다는 점으로부터 구절을 기본 단위로 이용하였으며, 연관 규칙 마이닝 방법 중 하나인 FP-Growth 기법을 활용하였다. 이로써 논문은 제안하는 프레임 워크를 통해 제조 설비의 오류 유형을 파악할 수 있고 설비 정지를 예방할 수 있으므로 설비 가동률을 높여서 생산성이 향상될 수 있다고 주장했다.

[3]은 많은 관측치 및 변수로 이루어진 대량의 제조 산업의 데이터에서 품질 문제 해결을 위한 통계적 프로세스 제어(SPC)를 수행하기 위해서 귀납적 학습과 인공신경망을 활용하는 방법을 제안하였다. 이를 위해 먼저 품질 관리에 필요한 변수를 줄이기 위한 특징 부분집합을 선택하는 방법을 제안하였다. 그 다음, 귀납적 학습의 정확한 예측률을 향상하기 위한 군집화와 귀납적 학습방법을 제시하였다. 그리고 끝으로 기준 패턴을 비교하여 서로 다른 패턴을 검출하기 위한 패턴 검출 방법을 제안하였다. 그리하여 연구에서 제안한 세 가지 방법이 연구에서 사용된 제작 데이터 셋과 같이 다 변량으로 구성된 데이터에 효과적인 것으로 나타났다.

[4]는 지능형 생산 시스템을 갖춘 위탁 생산 회사의 제조 관련 데이터를 수집하고 분석하는 시스템을 소개하고 텍스트 데이터를 분석하여 품질 불량에 영향을 미치는 핵심공정 또는 공정 패턴을 밝혀내는 실험을 진행하였다. 이를 통해 텍스트 마이닝 시스템의 성능을 향상하려면 도메인 및 상황에 맞는 적절한 튜닝 방법이 필요하다고 밝혔다.

[5]는 태양광 시스템을 가동할 때 나타나는 문제를 탐지하기 위한 통계 가설과 기계학습을 접목한 방법을 제시하였다. 이 연구에서는 GPR(Gaussian Process Regression)으로 모델링하고 GLRT(Generalized Likelihood Ratio Test) 방법으로 가상 및 실제 가동 데이터의 문제를 탐지하고 계산 시간과 허위 경보 비율을 측정하였다. 실험결과, 두 결과값이 서로 trade-off 관계로 밝혀졌다.

[6]은 호주의 발전 기업과 설탕 제조 기업에서 기록된 실제 설비정비의뢰서 및 설비가동중지 데이터의 텍스트 데이터를 이용하여 고장에 의한 설비의 가동중지시간을 예측하였다. 분류 알고리즘으로는 NB 와 SVM 두 가지를 각각 사용하였으며, 분류 결과 만족할 만한 결과를 얻었다.

[7]은 기계학습의 광범위한 영역과 제조업 분야의 복잡함으로 인해 제조업에 기계학습을 적용하기 어려운 점을 극복할 수 있도록 다양한 선행 연구의 개요를 정리하였다. 또한, 제조업 데이터의 경우 대부분 라벨 처리가 되어 있어 비지도 학습방법보다는 지도학습방법이 적합한 경우가 많고, 전문가의 지식을 활용할 수 있다는 점에서 논문을 작성할 당시에는 강화학습이 불필요하다고 주장하였다. 하지만, 가까운 미래에 데이터의 규모가 급속히 증가하게 되면, 강화학습의 수요도 증가할 것으로 예상했다.

2.2 자연어 처리(전처리, 임베딩, 단어 차원 축소) 관련 연구

[8]은 뉴스 텍스트 마이닝을 수행하여 뉴스가 주가에 호재 또는 악재일지 여부를 학습하고, 이를 바탕으로 새로 발행된 뉴스가 주가에 어떠한 영향을 미칠지를 예측하는 알고리즘을 제안하였다. 연구는 먼저 기존의 연구들이 제한된 양의 학습 데이터를 사용하고 특정 종목에

치우치는 문제점이 있다고 지적하면서 이를 극복하기 위해 4 년간의 다양한 종목군의 뉴스 데이터를 확보하였음을 밝혔다. 특성 추출은 성능이 우수한 것으로 알려진 BoW 알고리즘을 사용하였는데, 전처리 단계에서 단음절 단어 및 숫자를 제거하고 3 회 이상 등장하는 단어만 BoW 에 담았다. 그 다음으로 긍정 또는 부정으로 라벨을 부여하기 위해서 뉴스 발행 후 주가가 2% 이상 상승하면 긍정, 2% 이상 하락하면 부정으로 라벨을 부여하고 그 외의 경우는 학습데이터에서 제외시켰다. 이렇게 2005 년 1 월부터 2008 년 12 월까지 4 년간의 뉴스 데이터 중에서 조건에 부합하는 673 개 종목의 뉴스 42,355 건을 학습시킨 다음 2009 년 11 월 1 일부터 2010 년 2 월 28 일까지 4 개월간 발행된 뉴스로 예측한 결과 55.01%의 예측 성공률을 얻었다.

[9]는 문서 분류를 해결하는데 새로운 신경망 모델을 디자인하거나 파라미터를 최적화하는 것뿐만 아니라 워드 임베딩 모델과 전처리를 설계하는 것 또한 중요하다고 주장하고 적합한 전처리와 워드 임베딩의 조합을 찾는 방법을 제시하였다. 실험 데이터는 공개적으로 배포된 Zhang 등(2015)의 AG'S News 데이터셋을 사용하였고 알고리즘으로는 K-CNN 모델, Y-RNN 모델 그리고 L-RCNN 모델을 사용하였으며 임베딩 방법으로는 Skip-gram, GloVe 그리고 FastText 를 사용하였다. 전처리는 Lowering, Stemming, Lemmatizing, Punctuation-merging 그리고 Punctuation-spliting 을 조합하였고 Stop-words filtering 은 모든 경우의 수에 적용하였다. 실험은 총 세가지로 나뉘는데 먼저 대표 텍스트 전처리 타입 선정이다. 상기 텍스트 전처리 방법의 조합에 의해 만들어진 경우의 수에서 최종적으로 Type F, K, N, Q 네 가지 방법이 선택되었다. 다음 실험은 패딩과 미세조정 비교실험이다. 세 가지 알고리즘과 세 가지 임베딩의 조합에서 유의미한 패턴은 나타나지 않았다. 끝으로 세 번째 실험은 알고리즘, 임베딩 그리고 위에서 선택된 네가지 전처리 방법의 조합으로 이루어진 모든 경우의 수 별로 성능을 측정한 것이다. 그 결과 알고리즘 별로 결과를 봤을 때 K-CNN 과 Y-RNN 에서는 GloVe.840B 와 Type K 의 조합에서 91.733 / 92.224 와 92.104 / 92.461 의 가장 높은 성능을 보여주었고, L-RCNN 에서는 GloVe.840B 와 Type Q 에서 93.034 / 93.276 으로 성능이 가장 높았다.

[10]은 한국어와 중국어의 언어학적 특징을 고려한 문서 자동분류 시스템의 성능을 향상시킬 수 있는 자질어 단위를 제안하였다. 언어 종속적인 형태소 자질어와 언어 독립적인 n-gram 그리고 둘을 조합한 복합 자질어 집합을 구성하여 각 언어의 신문기사를 SVM 으로 분류하였다. 한국어 실험데이터는 6 개의 카테고리의 24,605 개의 한국어 인터넷 신문기사를 사용하였고, 테스트 데이터로 각 카테고리당 350 개씩 추출하여 총 2,100 개의 신문기사를 사용하였다. 중국어 실험 데이터는 8 개 카테고리의 신문기사 총 20,127 개의 신문기사를 사용하였고, 테스트 데이터로 각 카테고리에서 300 개씩 총 2,400 개의 신문기사를 사용하였다. 결론부터 언급하면 한국어 문서분류에서는 Bigram 이 F1-measure 87.07%로 가장 좋은 분류 성능을 보였고, 중국어 문서분류에서는 'Unigram-명사-동사-형용사-사자성어'의 복합 자질어 집합이 F1-measure 82.79%로 가장 좋은 성능을 보였다. 논문에서는 그 이유를 언어 종속적인 특징인 형태소보다 언어 독립적인 특징인 N-gram 이 성능향상에 좀 더기여한 것으로 보았다. 그리고 실험 데이터에서 해당 N-gram 으로 표현 가능한 특징들이 충분히 표현되어야 하므로 N 을 정하는데 신중을 기해야 한다고 주장하였다. 그 외 Unigram 과 Bigram 의 합성인 누적 Bigram 의 경우는 영어를 포함한 많은 언어에서 높은 성능을 보여주므로 가장 보편적으로 사용될 수 있다고 주장하였다.

[11]은 TF-IDF 로 문서를 벡터화하고 Singular Value Decomposition 으로 단어 차원을 축소한 다음 K-means 로 클러스터링하는 하이브리드 알고리즘을 제시하며, BBC news 와 BBC sport 기사들을 주제별로 나누는 실험을 진행하였다. 실험결과, 제시된 하이브리드 알고리즘이 단어 차원의 축소에 매우 효과적임에도 클러스터링의 정확도에 영향을 거의 주지 않는 것으로 나타났다.

[12]는 트위터와 같이 노이즈 텍스트 데이터를 분석할 때는 전처리 단계가 매우 중요함을 밝히며, 노이즈 텍스트를 위한 2 단계의 전처리 방법을 제시하였다. 또한, 건강 관련 트위터 텍스트 데이터의 본래 의도를 분류한 실험을 진행해, 전처리하지 않은 것 대비 5.4%가량 성능이 향상된 결과를 보여주었다.

[13]은 전처리 방법의 선택이 결과에 미치는 영향을 조사하기 위해 다양한 분야의 텍스트 데이터에 Vanilla, Lemmatizing, Lower-casing 그리고 Multi-word grouping 전처리 방법을

적용하여 토픽 모델링과 극성 분류 실험을 진행하였다. 실험결과, 데이터가 충분할 경우 일반적인 텍스트 데이터에서는 가장 간단한 방법인 Vanilla 가 오히려 우수하면서도 고른 성능을 보이는 것으로 나타났다. 이에 비해, 의료 분야와 같이 특수한 도메인의 텍스트 데이터에서는 고유 명사 처리 방법인 Multi-word grouping 을 적용했을 때 눈에 띄게 성능이 향상된 것으로 관찰되었다. 저자는 그 원인이 빈번한 전문용어 및 고유 명사의 출현에 있다고 지목했다. 덧붙여, 텍스트 데이터의 양이 제한적일 경우에는 Vanilla 외 여러 방법이 충분히 고려될 수 있다고 하였다.

III. 연구 방법

본 연구는 원자력발전소에 설치되는 기기들을 제작하는 공장에서 실제 6 년 동안 기록된 제작 기록의 텍스트 데이터와 품질 불량 기록 데이터를 이용하여 그림 2 와 같이 품질 불량 위험 공정을 예측하는 실험을 진행하였다.

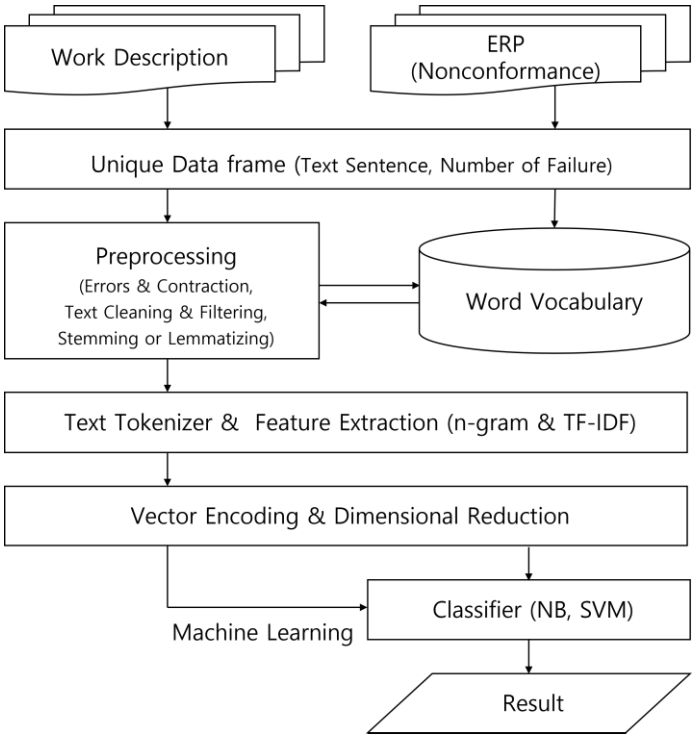

Figure 2. Flow chart of the study
그림 2. 연구 흐름도

3.1 데이터 수집 및 전처리 과정

원자력발전소 기기 제작 프로젝트에서는 미국기계기술자협회(ASME) 또는 한국전력산업기술기준(KEPIC) 코드에서 규제하는 원자력 기기 설계 및 제작검사 기준에 따라 이를 제작해야 한다. 또한, 추적관리 및 안전보증을 위해 모든 제작 과정을 기록한다. 트래블러는 제작 요건에 따라 공정을 설계한 문서로, 여기에는 모든 제작 관련 내용이 기록되어 있다. 본 논문은 전산화된 트레블러 상의 제작공정을 설명하는 문장 텍스트 데이터 53,863 개를 독립변수로 사용하였고, 각각을 품질 불량 발생 여부에 따라 1 과 0 으로 라벨을 부여하였다. 여기서 통계적 유의성을 확보하고자 수행 횟수가 10 회 미만인 공정들을 제외하기 위하여

공정을 설명하는 텍스트를 기준으로 유니크 단위로 grouping 한 결과, 표 1 과 같이 1,174 종류의 공정 그룹별로 라벨이 부여된 유니크 데이터 프레임이 생성되었다.

전체적인 전처리의 순서는 다음과 같다. 우선 오탈자 및 띄어쓰기 오류를 교정하기 위해 MS word 의 교정 추천 기능을 활용하였다. 그 다음으로 전처리 과정에 도메인 지식을 반영하기 위한 핵심인 축약어 처리를 수행하였다. 원자력발전소 주기기의 제작 문서 내 텍스트는 수십 년간 특정 도메인 영역 안에서 특정 상황들을 묘사하는 데 사용된 관계로, 뉴스 등에 등장하는 일반 단어의 쓰임과는 차이가 있다. 표 2 와 같이 다양한 축약 방식이 활용되는 경우가 빈번한데, 'C/HEAD'와 'T/A'는 각각 'CHANNEL HEAD'와 'TEMPORARY ATTACHMENT'의 축약 표현이다. 따라서 도메인 지식을 기반으로 단어 사전을 생성하고 이를 맵핑하여 축약어를 원래의 단어로 치환하는 방식이 전처리 과정에서 도메인 지식을 반영하고 잡음을 줄이는 데 효과적이라고 할 수 있다. 중요한 점은 이 단계가 기호 제거 이전에 수행되어야 한다는 것인데, [14]와 같이 기호가 제거된 상태에서는 축약어에 해당하는지 구별할 수 없기 때문이다.

Table 1. Unique operation data frame
표 1. 유니크 공정 데이터 프레임

ROW	DESCRIPTION	NUMBER OF EXECUTIONS	NUMBER OF FAILURE	LABEL
1	U.T. ON C/HEAD G/SEAM AREA	78	9	1
2	LOCAL PWHT	87	0	0
3	M/C 2-W.P.	35	3	1
⋮	⋮	⋮	⋮	⋮
1173	FIT-UP D.C.F.W P/P	13	0	0
1174	REMOVAL T/A ON 3RD HALF EGC(IF NECESSARY)	55	0	0

Table 2. Examples of preprocessing
표 2. 전처리 전-후 예시

ROW	BEFORE PREPROCESSING	AFTER PREPROCESSING
1	U.T. ON C/HEAD G/SEAM AREA	ULTRASONIC TEST CHANNEL HEAD GIRTH SEAM AREA
2	LOCAL PWHT	LOCAL POST WELD HEAT TREATMENT
3	M/C 2-W.P.	MACHINE WELD PREPARATION
⋮	⋮	⋮
1173	FIT-UP D.C.F.W P/P	FIT UP
1174	REMOVAL T/A ON 3RD HALF EGC(IF NECESSARY)	REMOVE TEMPORARY ATTACHMENT THIRD EGGCREATE

다음으로 모든 텍스트가 영어 대문자로 쓰였기 때문에 Lower-casing 은 생략하였고, 프로그래밍 언어인 R 로 읽어올 때 줄 바꿈 기호 '\n'이 표기되어서 이를 제거하였다. 그리고 Trimming 으로 앞뒤 공백을 제거하고, 두 칸 이상 띄어쓰기가 되어 있는 White-space 를 한 칸의 띄어쓰기로 변환시켰다.

끝으로 선택적 Stopwords filtering 과 선택적 Stemming or Lemmatizing 단계를 거쳤다. 일반적으로 전치사, 관계사, 접속사는 Stopwords 에 속하지만 본 연구에서는 in, out, within, without, before 그리고 after 와 같이 특정 조건을 나타내는 것들은 보존하였다. Stemming or Lemmatizing 단계 또한 중요한데, 일반적인 것을 적용하면 원래의 고유 명사가 명사나 동사로 쓰이는 경우와 구분되지 않는 경우가 빈번하기 때문이다. 예를 들어 'SEPARATOR SUPPORT'라는 하나의 고유 명사에 일반적인 Stemming 을 적용하면 'SEPARAT SUPPORT'로 변환되어 각각 명사나 동사로 쓰이는 경우와 구분되지 않는 문제가 생긴다. 이렇게 하여 최종적으로 유니크 공정 설명문장의 수는 1,081 개가 되었다.

Table 3. Example of N-gram (accumulative Trigram) and TF-IDF
표 3. N-gram(누적 Trigram) 및 TF-IDF 적용 예시

N-grams	TEXT_1	TEXT_2	TEXT_3	TEXT_4	TEXT_5	TEXT_6	TEXT_7	TEXT_8	TEXT_9	TEXT_10	. . .
adjust	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
divider	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
plate	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
section	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
adjust_divider	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
divider_plate	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
plate_section	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
adjust_divider_plate	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
divider_plate_section	**0.1000**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
assemble	0.0000	**1.0000**	**0.1333**	**0.1667**	**0.1667**	**0.1000**	**0.0333**	**0.1667**	**0.1000**	**0.0500**	. . .
channel	0.0000	0.0000	**0.1333**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
shroud	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
assemble_channel	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
channel_channel	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
channel_shroud	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
shroud_assemble	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
assemble_channel_channel	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
channel_channel_shroud	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
channel_shroud_assemble	0.0000	0.0000	**0.0667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
dryer	0.0000	0.0000	0.0000	**0.1667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
vane	0.0000	0.0000	0.0000	**0.1667**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	. . .
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.

Table 4. Example of SVD
표 4. SVD 적용 예시

TEXT	V_1	V_2	V_3	. . .	V_298	V_299	V_300	LABEL
TEXT_1	-0.00474358	0.01048510	0.00027838	. . .	-0.00084538	-0.01114794	-0.06719222	0
TEXT_2	-0.00545804	0.01123521	0.00235655	. . .	-0.00293463	-0.00237903	0.00167368	0
TEXT_3	-0.01212619	0.02541402	0.00508957	. . .	0.11478770	-0.06162550	0.05488070	0
TEXT_4	-0.00592381	0.01262663	0.00282878	. . .	-0.01454897	0.00959943	0.00036770	0
TEXT_5	-0.03277413	0.05707330	-0.03503337	. . .	0.00741308	-0.00019424	-0.00752750	0
TEXT_6	-0.01337963	0.01197552	0.00561431	. . .	0.01427222	0.03405350	0.03791498	1
TEXT_7	-0.00723548	0.02275435	0.00723242	. . .	-0.00053434	-0.00745825	0.00690870	0
TEXT_8	-0.01091755	0.01263778	0.01086243	. . .	-0.00357309	-0.00514138	-0.00152443	0
TEXT_9	-0.00559995	0.01573534	0.02605127	. . .	-0.00447045	-0.00667166	-0.00368653	0
TEXT_10	-0.00646905	0.01178007	0.00247528	. . .	0.01562638	0.01627170	-0.01114794	1
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.

3.2 문장 임베딩 및 교차 검증

N-gram 및 TF-IDF 로 문장 텍스트의 특징을 추출한 다음 SVD(Singular Value Decomposition)을 활용하여 차원을 축소하였다. N-gram 은 단어의 출현 순서를 반영하기 때문에 각각의 단위가 상대적인 의미를 지닐 수 있게 해준다. 본 연구에서는 데이터의 크기와 실험 환경을 고려하여 Unigram 부터 누적 Trigram 까지 적용하였다. 표 3 은 누적 Trigram 의 예인데 Unigram, Bigram 그리고 Trigram 의 예시를 나타낸다. TF-IDF 는 단어의 개별 의미가 문장의 전체 의미에 이바지하는 정도를 가중치로 나타내는 데 쓰인다. 가령, 1) CHANNEL HEAD GIRTH SEAM WELDING 2) FINAL VESSEL GIRTH SEAM WELDING 3) CHANNEL HEAD FINAL MACHINING 의 세 공정 설명 중에 1 과 2 의 관계가 1 과 3 의 관계보다 더 가깝다면, 각각 공유된 단어인 ‘GIRTH’, ‘SEAM’, ‘WELDING’이 ‘CHANNEL’, ‘HEAD’보다 문장 전체의 의미에 이바지하는 정도가 더 크다고 할 수 있다.

SVD 의 차원은 [15]에서 제시한 대로 300 으로 정하였다. 그 다음 교차 검증을 위해 142 개의 1(13.14%)과 그 외 0(86.86%)으로 이루어진 1,081 개의 행 데이터를 5-fold 로 나눴다.

IV. 실험 결과

4.1 텍스트 전처리 방법 및 n-gram 별 분류 성능

기계학습 분류 알고리즘으로는 분류 문제에서 일반적으로 사용되고 있는 NB(Naive Bayes classifier)와 SVM(Support Vector Machine) 두 가지를 사용하여 실험의 객관성을 높였다. 두 알고리즘 모두 R 의 e1071 패키지를 활용하였는데, 파라미터들은 SVM 의 분류 타입을 'C-classification'으로, Kernel 을 'Linear'로 지정한 것 이외에는 모두 e1071 패키지의 디폴트 값을 따랐다. 그리고 본 연구에서 제안한 전처리 과정의 효과를 확인하기 위해 전처리를 하지 않은 상태 및 일반적인 전처리를 수행한 상태와 비교하였다. 일반적인 전처리는 R 의 'tm' Package 를 활용하여 Numbers, Punctuation, White-space, Stopwords 그리고 Stemming 을 수행한 것을 말한다. 또한 Unigram, Bigram 그리고 Trigram 으로 케이스를 나눠 N-gram 별 변화도 살펴보았다.

열별로 높은 성능을 보여준 것을 굵은 글씨로 표시하였다. 실험결과를 종합해보면 SVM 이 NB 보다 확실히 높은 성능을 보여주었는데, 이는 NB 가 문장 텍스트와 같이 고차원의 벡터를 분류하는 데 적합하지 않은 특징 때문으로 해석된다. 그리고 본 연구에서 제안한 전처리 방식이 그 외의 경우와 비교하여 전체적으로 높은 성능을 보여주기 때문에 본 연구에서 제안한 전처리 방식이 도메인 지식을 잘 반영했다고 볼 수 있다. 그 밖에 Unigram 에서 Trigram 으로 갈수록 대체로 좋은 결과를 보여주었다.

Table 5. Result according to preprocess type, classifier and n-gram
표 5. 전처리 방식, 분류기 그리고 n-gram 에 따른 분류결과

preprocessing	classifier	uni-gram		bi-gram		tri-gram	
		acc	f1	acc	f1	acc	f1
none	NB	0.5724	0.6931	0.6244	0.7409	0.6044	0.7282
	SVM	0.7043	0.8050	0.7232	0.8313	0.7262	0.8332
general	NB	0.5624	0.6867	0.6024	0.7238	0.6134	0.7353
	SVM	0.6962	0.8008	0.7213	0.8300	0.7472	0.8480
proposed	NB	0.5924	0.7367	0.6244	0.7479	0.6314	0.7489
	SVM	**0.7051**	**0.8192**	**0.7361**	**0.8438**	**0.7685**	**0.8641**

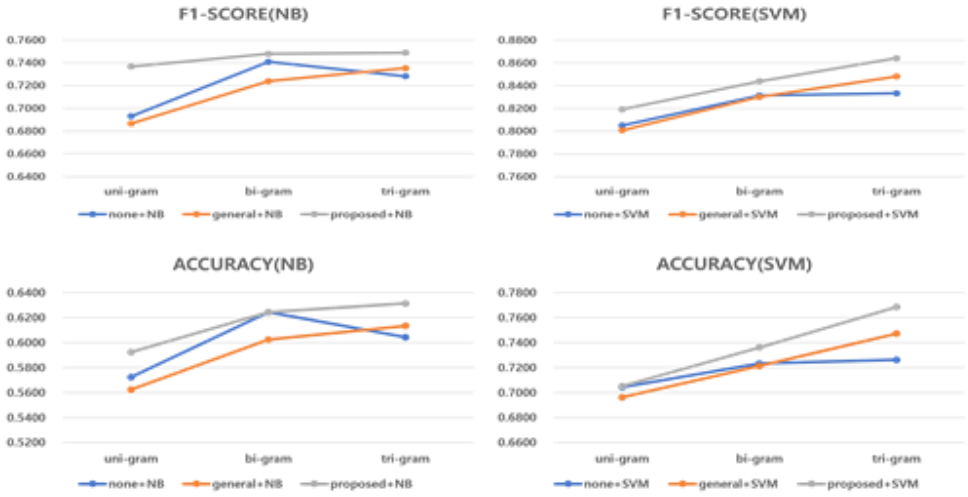

Figure 3. F1-scores and accuracies
그림 3. 각 F1-score 및 accuracy

그 결과 전반적으로 현장 엔지니어들의 예측과 알고리즘으로 분류한 결과가 동조하는 경향을 보였다. 즉, 알고리즘이 품질 불량 발생 가능성이 있다고 판단하여 1 로 라벨링한 공정들은 엔지니어들의 투표수 또한 높았다. 구체적으로 살펴보면 분류기는 40 번, 50 번, 70 번, 80 번, 120 번 그리고 130 번 공정을 1 로 라벨링 하였는데 엔지니어들의 투표수가 비교적 높은 공정들이다.

Table 6. Evaluation between prediction and professional's decision
표 6. 새로운 제작공정에 대한 예측과 전문가 집단 투표 간 비교

OPER NO.	OPERATION DESCRIPTION	PREDICTION	EXPERTS' VOTING
10	CHECK APPROVED DOCUMENT FOR MAINTAINING ON SHOP	0	
20	REVIEW OF WELDER QUALIFICATION & VERIFICATION OF WELDING PROCEDURE SPECIFICATION TO BE USED	0	
30	VERIFICATION OF MATERIAL IDENTIFICATION AND CERTIFICATION	0	
40	THIMBLE GUIDE TUBE BENDING	1	
50	VISUAL & DIMENSION INSPECTION OF THIMBLE GUIDE TUBE	1	
60	PREPARATION FOR HYDROSTATIC TEST THIMBLE GUIDE TUBE	0	
70	NDE-PT EXAMINATION ON END OF THIMBLE GUIDE TUBE	1	
80	HYDROSTATIC TEST THIMBLE GUIDE TUBE	1	
90	CUTTING OF THIMBLE GUIDE TUBE	1	
100	VISUAL & DIMENSION INSPECTION(BALL TEST)	0	
110	NDE-UT EXAMINATION(UT THIKNESS) ON THIMBLE GUIDE TUBE	1	
120	MACHINING OF THIMBLE GUIDE TUBE	1	
130	VISUAL & DIMENSION INSPECTION	1	
140	ISSUED CMTR/	0	
150	CLEANING THIMBLE GUIDE TUBE	0	
160	CLEANING INSPECTION THIMBLE GUIDE TUBE	0	
170	QA DOCUMENT PACKAGE REVIEW	0	
180	PACKING _ THIMBLE GUIDE TUBE	0	
190	PACKING INSPECTION _ THIMBLE GUIDE TUBE	0	
200	SHIPPING _ THIMBLE GUIDE TUBE	0	
210	SHIPPING INSPECTION _ THIMBLE GUIDE TUBE	0	

V. 결론 및 논의

본 연구는 실제 제조업 현장에서 발생한 데이터를 자연어 처리 및 기계학습을 활용하여 품질 불량 위험 공정을 예측하는 방법을 제안하였다. 전처리 단계에서는 도메인 지식이 잘 반영될 수 있도록 단어 사전을 활용하여 오탈자, 축약어, 불용어를 처리하고 Stemming 을 수행하는 과정을 설계하였고, 이를 실험을 통해 증명하였다. 그리고 문장 벡터화 단계에서는 N-gram 및 TF-IDF 와 SVD 를 결합한 하이브리드 알고리즘을 제시하였다. 이것을 NB 와 SVM 으로 분류한 결과 누적 Trigram 을 적용하고 SVM 을 사용하였을 경우 Accuracy 와 F1-score 가 각각 0.7685 와 0.8641 로써 본 연구에서 제안한 방법이 상당히 유효할 수 있음을 보여주었다. 여기에 덧붙여 새로운 공정에 대한 예측과 현장 엔지니어들의 투표 간의 비교를 통해 실제 현장에서 자연스럽게 활용될 수 있음을 확인하였다.

반면에 본 연구는 품질 불량 위험 공정을 예측하면서 작업 설명 텍스트라는 내적 요인만을 변수로 활용하고 작업자의 기량, 제작 부하 등과 같은 외적 요인들은 고려하지 않았다는 점에서 한계가 있다. 품질 불량 발생 가능성은 주변의 여러 환경적 요소에 따라 달라질 수 있기 때문이다. 하지만 이 문제는 제조업 현장의 데이터 마이닝 수준이 더욱 고도화된다면 충분히 해결될 수 있을 것으로 기대한다. 그리고 품질 불량 발생 가능성을 예측한 본 연구에 이은 향후 연구로는 품질 불량이 발생했을 경우에 생길 수 있는 대기기간, 비용 등을 예측하는 연구가 적합할 것으로 생각된다.

VI. Acknowledgment

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2020R1G1A1099559).

VII. 참고문헌

[1] B. Park and W. Lee. “Use Cases of Machine Learning Techniques for Manufacturing Process Data: Comparative Analysis of CART, Random Forest, and TreeNet,” Quality Management of Korea, 2021(0), pp. 40-40, 2021

[2] J. Yoon, H. An and Y. Choi. “A Machine Learning Based Facility Error Pattern Extraction Framework for Smart Manufacturing,” Society for E-Business Studies 23(2), pp. 97-110, 2018

[3] B. Kang, S. Park. “Integrated machine learning approaches for complementing statistical process control procedures.” Decision Support Systems, 29(1), pp. 59-72, 2000

[4] G. Li, Y. Cao, S. Zhao, Y. Bao and C. Yu. “Research on Text Data Pool of Intelligent Manufacturing for Plate Parts.” Journal of Physics: Conference Series, 1575(012199), 2020

[5] R. Fazai, K. Abodayeh, M. Mansouri, M. Trabelsi, H. Nounou, M. Nounou and G.E. Georghiou. “Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems.” Solar Energy, 190, pp. 405-413, 2019

[6] K. Arif-Uz-Zaman, M. E. Cholette, L. Ma and A. Karim. “Extracting failure time data from industrial maintenance records using text mining.” Advanced Engineering Informatics, 33, pp. 388-396, 2017

[7] T. Wuest, D. Weimer, C. Irgens and K. D. Thoben. “Machine learning in manufacturing: advantages, challenges, and applications.” Production & Manufacturing Research, 4(1), pp. 23-45, 2016

[8] S. An and S. Jo. “Stock Prediction Using News Text Mining and Time Series Analysis,” Korea Computer Congress 2010, 37(1), pp. 364-369, 2010

[9] Y. Kim and S. Lee. “Combinations of Text Preprocessing and Word Embedding Suitable for Neural Network Models for Document Classification,” Journal of KIISE, 45(7), pp. 690-700, 2018

[10] M. Lim and S. Kang. “Comparison Between Optimal Features of Korean and Chinese for Text Classification,” International Journal of Fuzzy Logic and Intelligent Systems, 25(4), pp. 386-391, 2015

[11] A. I. Kadhim, Y. Cheah and N. H. Ahamed. “Text Document preprocessing and Dimension Reduction Techniques for Text Document Clustering.” 2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology, pp. 69-73, 2014

[12] C. S. Pavan Kumar and L. D. Dhinesh Babu. “Novel Text Preprocessing Framework for Sentiment Analysis. Smart Intelligent Computing and Applications. Smart Innovation,” Systems and Technologies, pp. 105, 2019

[13] J. Camacho-Collados and M. T. Pilehvar. Cornell University (Ver. 3) https://arxiv.org/abs/1707.01780v3 (downloaded:2020. 9. 20)

[14] J. Perkins. Python 3 Text Processing with NLTK 3 Cookbook, 2nd Ed., pp. 36. (Packt Publishing Ltd., Birmingham)

[15] Jing Gao and Jun Zhang. “Clustered SVD strategies in latent semantic indexing,” Information Processing & Management 41(5), pp. 1051-1063, 2005

저자 소개

노정민*(Jeong-Min Roh)*

2019 년 3 월~현재 고려사이버대학교 석사과정
2008 년 10 월~현재 두산중공업 재직 중

관심분야 : 인공지능, 제조업

김용성*(Yongsung Kim)*

2013 년 8 월 고려대학교 전자컴퓨터공학 석사
2018 년 8 월 고려대학교 컴퓨터공학 박사
2018 년 8 월~2020 년 2 월 소프트웨어정책연구소 선임연구원
2020 년 3 월~현재 고려사이버대학교 소프트웨어공학과 조교수

관심분야 : 인공지능, 머신러닝, 학습분석, AI/SW 교육, AI/SW 인재양성