

Classifying Sub-Categories of Apartment Defect Repair Tasks: A Machine Learning Approach

Eunhye Kim[†] · HongGeun Ji^{††} · Jina Kim^{†††} · Eunil Park^{††††} · Jay Y. Ohm^{†††††}

ABSTRACT

A number of construction companies in Korea invest considerable human and financial resources to construct a system for managing apartment defect data and for categorizing repair tasks. Thus, this study proposes machine learning models to automatically classify defect complaint text-data into one of the sub categories of 'finishing work' (i.e., one of the defect repair tasks). In the proposed models, we employed two word representation methods (Bag-of-words, Term Frequency-Inverse Document Frequency (TF-IDF)) and two machine learning classifiers (Support Vector Machine, Random Forest). In particular, we conducted both binary- and multi- classification tasks to classify 9 sub categories of finishing work: home appliance installation work, paperwork, painting work, plastering work, interior masonry work, plaster finishing work, indoor furniture installation work, kitchen facility installation work, and tiling work. The machine learning classifiers using the TF-IDF representation method and Random Forest classification achieved more than 90% accuracy, precision, recall, and F1 score. We shed light on the possibility of constructing automated defect classification systems based on the proposed machine learning models.

Keywords : Apartment, Defect, Repair Tasks, Sub Category, Finishing Works, Machine Learning

아파트 하자 보수 시설공사 세부공종 머신러닝 분류 시스템에 관한 연구

김은혜[†] · 지홍근^{††} · 김지나^{†††} · 박은일^{††††} · 옴재용^{†††††}

요 약

대한민국 건설사들은 아파트 하자 정보를 추적하고 보수작업을 관리하기 위한 시스템을 운영하는데 상당한 인력과 비용을 투자하고 있다. 본 연구에서는 하자 접수 상세내용 텍스트 데이터를 이용하여 하자 보수 시설공사에 따른 세부공종을 분류하는 머신러닝 모델을 제안한다. 두 가지 단어 임베딩(Bag-of-words, Term Frequency-Inverse Document Frequency (TF-IDF))과 두 가지 분류기(Support Vector Machine, Random Forest)를 통해 한국어로 작성된 65만건 이상의 하자 접수데이터로부터 하자보수 시설공사 세부공종을 분류했다. 특히, 이번 연구에서는 특정 시설공사(마감공사)의 9개 세부공종(가전제품, 도배공사, 도장공사, 미장공사, 석공사, 수장공사, 옥내가구공사, 주방기구공사, 타일공사)을 분류하는 이진 분류 모델과 다중 분류 모델을 연구했다. 그 결과, TF-IDF와 Random Forest를 사용한 두가지 분류 모델에서 90%이상의 정확도, 정밀도, 재현율 및 F1점수를 확인했다.

키워드 : 아파트, 하자, 시설공사, 세부공종, 마감공사, 머신러닝

1. 서 론

전 세계적으로 도시 지역의 인구 폭증과 함께, 주요 국가의 도시 내 주거 유형 트렌드는 급변하고 있다. 그 중에서도

대한민국은 2019년 대한민국 주거 유형의 절반 이상이 아파트인 것으로 집계되었다(Table 1)[1].

통상적으로 아파트란 “독립된 공간을 소유한 개인 세대가 모여서 거주하는 5층 이상의 건축물”을 의미한다[2]. 이러한 아파트 수요 급증에 발맞추어 대한민국을 포함하여 인구 밀집도가 높은 일부 국가에서는 아파트형 주택을 주요 거주 형태 중 하나로 채택하였으며, 이 과정에서 원활한 주택 공급을 가능하게 하는 선분양 제도를 채택하였다.

선분양 제도란 “착공 시점에 입주인이 계약금을 지불하고 준공 시점에 잔금을 지급하여 주택에 입주하는 제도”이다[3]. 입주인 입장에서 선분양 제도는 준공 시점의 분양가를 토대로 계약금이 산정되고 잔금을 지불하기까지의 일정 시간이 주어 진다는 경제적인 혜택이 있지만 많은 단점 또한 존재한다. 특

※ 이 논문은 과학기술정보통신부/정보통신기획평가원의 ICT혁신인재4.0 사업(IITP-2020-0-01816)과 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2019R111A2A01058640).

† 비 회 원 : 성균관대학교 데이터사이언스융합학과 석사과정

†† 준 회 원 : 성균관대학교 인공지능융합학과 석사과정

††† 비 회 원 : 성균관대학교 인터랙션사이언스 석사과정

†††† 비 회 원 : 성균관대학교 인공지능융합학과 조교수

††††† 비 회 원 : 한국과학기술원 기술경영전문대학원 교수

Manuscript Received : February 2, 2021

First Revision : April 22, 2021

Accepted : May 1, 2021

* Corresponding Author : Eunil Park(eunilpark@skku.edu)

Table 1. Residential Type in South Korea, 2019

Residential Type	Ratio[%]
Apartment	50.1
Detached housing	32.1
Multiplex housing	9.4
Small apartment	2.2
House within commercial building	1.6
Living quarter other than housing unit	4.6
Total	100

히, 준공 상태의 아파트 상태와 품질을 직접 확인하고 주택을 구입할 수 없기 때문에 품질 확보 및 고객 만족과 관련 하여 입주민과 시공사 간에 상당한 분쟁이 발생하고 있다[3].

이 중, 아파트 입주민과 시공사는 건축물 하자 발생 및 보수 작업과 관련하여 크고 작은 분쟁을 겪고 있다. ‘하자’는 안전상, 기능상, 미관상의 지장을 초래하는 건축물 및 시설의 결함을 의미한다[2]. 이러한 하자관련 분쟁과 소송을 조정하기 위해 대한민국 정부는 ‘하자심사·분쟁조정위원회’를 설립하고 급증하는 하자 소송과 분쟁을 효율적이고 효과적으로 조정하기 위한 노력을 기울이고 있다[3]. Fig. 1과 같이 ‘하자심사·분쟁조정위원회’에 접수된 분쟁 기록에 따르면 2010년 69건으로 집계되었던 공동주택 하자 분쟁건수가 2019년에는 약 62배 증가한 4,290건이 집계되었다[4].

또한, 국토교통부(주택건설공급과)는 ‘공동주택 하자의 조사, 보수비용 산정 및 하자판정기준’을 마련하고 23개의 시설공사별 세부공사(마감, 창호, 급·배수 및 위생설비 등)를 분류하고 이에 따른 담보책임기간을 지정했다[5].

대부분의 대한민국 대형 건설사들은 고유한 사내 하자관리 시스템을 활용하여 하자 데이터를 수집하고 있다(Fig. 2). 본 연구에서 활용된 데이터 수집 대상의 건설사에서는 입주민이 건축물 하자를 발견하면 고객 지원 센터에 통화, 휴대폰 애플리케이션, 방문 등의 방법으로 하자를 접수한다. 그 후, 접수된 내용을 토대로 고객 지원 센터 상담원이 일련의 하자 기본 정보(접수일, 하자 위치, 세대 동/호수, 하자 유형, 하자 예상 원인, 하자 상세 내용)를 사내 하자관리 시스템에 입력하고 시설공사 및 세부공종을 선택한다. 이때 선택된 작업공종에 따라 전문 업체에 해당 작업 지시가 전달되고 자재 확보가 이

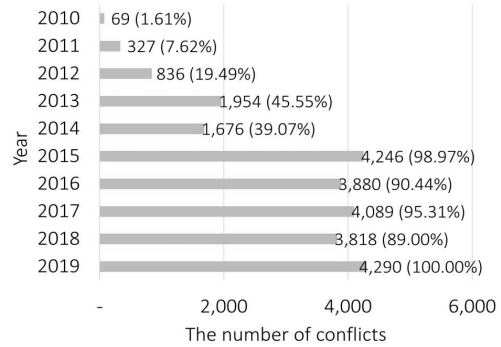


Fig. 1. The Number of Apartment Defect Conflicts in South Korea

루어진다. 그렇기 때문에 정확하고 신속한 작업공종 선택은 추후 작업 관리 및 품질 확보에 큰 영향을 주는 중요한 작업이라고 여겨진다.

이에 본 연구에서는 아파트 입주민 하자 접수 데이터 중, 마감공사를 필요로 하는 하자 접수 건을 대상으로 총 9개의 마감 세부공종(도배공사, 타일공사, 주방기구공사, 욕내기구공사, 도장공사, 수장공사, 미장공사, 석공사(건물내부공사), 가전제품)으로 분류하는 머신러닝 모델을 구축하고 평가하였다.

2장에서는 하자 보수 작업공종 분류에 관한 선행 연구를 소개하고, 3장에서는 머신러닝 모델관련 실험내용을 설명한다. 4장에서는 머신러닝 모델에 대한 평가를 제시하고 5장에서 본 연구의 결론과 시사점을 제시한다.

2. 선행 연구

2.1 아파트 하자 분류 방법

건설 산업에서는 작업 관리 효율화를 위하여 다양한 접근 방식으로 하자 유형 분류 관련 연구가 진행되었다. 먼저, Lee et al.[6]은 손실 분배 접근 방식을 사용하여 8 개의 건물 결함의 잠재적 위험을 추정했다. 또한, 위험 수준을 줄이기 위해 7 가지 유형의 보수공종 분류 방법을 제안했다 (예: 철근 콘크리트 (RC), 조적, 마감, 기계/전기/배관 (MEP), 창호, 가구 및 기타). 또한, Kim et al. [7]은 공동주택 하자의 보증 기간에 따른 패턴을 분석하기 위해, 작업 유형, 하자 유형 및 위치를 고려한 분류 시스템을 제안했다. 특히 마감 보수공사를 요하는 하자는 다양한 발생 원인, 불확실성, 위험 요소와

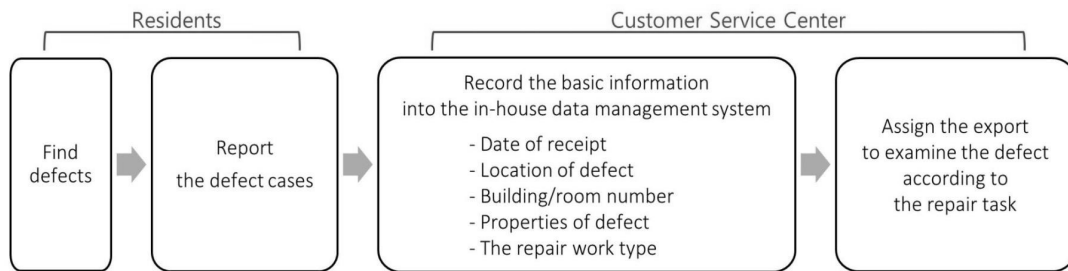


Fig. 2. Work Process

연관되어 있을 수 있어 보수 작업 및 유지 관리를 위해 더 많은 관심과 기술력이 필요하다[8]. 이와 관련하여 Park et al. [8]은 한국 아파트 단지에 대한 하자 보수 관련 자료를 수집하고 접수된 하자 건수와 마감 작업 빈도 사이에 유의한 연관성이 있음을 확인했다. 또한 연간 유지 보수 빈도 데이터를 분석하여 마감 작업의 각 구성 요소에 대한 유지 보수 계획의 패턴을 조사했다.

즉, 선행연구에 따르면 보수 작업과 마감공사에 유의한 상관관계가 있다는 연구 결과를 도출하였으며, 이를 바탕으로 본 연구에서는 입주민 하자 상세 내용을 통해 마감 공사 세부공종을 자동 분류하는 머신러닝 모델을 설계했다.

3. 실험 방법

본 연구의 전체 개요는 Fig. 3과 같다.

3.1 데이터 수집 (Data Collection)

본 연구에서는 2019년 국토교통부 시공능력평가 10위 이내 민간 건설업체의 19개 아파트 현장에서 수집된 입주민 하자 접수데이터 656,266건을 활용했다(Table 2). 해당 건설업체는 2019년에는 21,000세대 이상, 2020년에는 34,000세대이상의 아파트를 분양했으며 이에 따른 보수작업 품질확보에 많은 노력을 기울이고 있다. 이번 연구에서는 수집 대상 현장은 2020년 기준, 3년차와 4년차 단계의 아파트 현장으로 제한했다.

전체 하자 접수 건을 시설공사 세부공종 기준으로 나누어 보았을 때, Table 3와 같이 마감공사(Finishing work)가 56.85%로 가장 많은 비율을 차지하고 있다. 본 연구에서 분류하고자 하는 마감공사 세부공종은 Table 4와 같이 9개의 카테고리로 나누어져 있다. 이중, 가장 큰 비율을 차지하는 세부공종은 '도배공사(Painting work)'로 마감공사 중, 24.65%를 차지하고 있다.

3.2 데이터 전처리(Data Preprocessing)

'하자 상세내용' 텍스트 데이터 전처리를 위하여 아래의 절차를 진행하였다.

- ① 문장 부호, 추가 공백 문자, 단일 자음/모음을 제거
- ② 텍스트의 토큰화(Tokenization)를 진행
- ③ 텍스트의 어간(Stem)을 추출
- ④ NLTK2에서 제공하는 불용어(Stopword)를 제거
- ⑤ 문자열 길이가 10이상인 데이터만을 선별
- ⑥ 결측치를 제거

3.3 단어 임베딩(Word Embedding)

머신러닝 분류 모델(이중 분류, 다중 분류)을 훈련시키기 위해 전처리한 데이터를 두 가지 단어 임베딩 방법을 이용하여 벡터화 시켰다.

· **BoW(Bag of Words)**는 텍스트 데이터에서 각 토큰의 순서

Table 2. Overview of Data Collection

Site	Year of completion	Number of households	Number of defects
1	2018	1,730	112,326
2	2017	995	57,109
3	2018	1,164	53,744
4	2017	921	47,456
5	2017	1,077	42,717
6	2018	964	40,805
7	2018	913	36,986
8	2017	1,259	34,281
9	2018	1,160	27,895
10	2018	620	26,928
11	2018	975	26,520
12	2018	1,135	25,578
13	2017	832	23,496
14	2017	1,510	22,336
15	2018	629	21,859
16	2017	996	18,740
17	2018	677	14,332
18	2017	907	13,296
19	2017	759	9,862
Total		19,223	656,266

Table 3. Defect Frequencies for Each Repair Task

Repair task	Number of defects	Ratio [%]
Finishing work	373,118	56.85
Plumbing installation	73,612	11.22
Window work	73,226	11.16
Electricity and power facility installation	39,628	6.04
Heating, ventilating, cooling, air conditioning equipment installation	28,199	4.30
Wood work	25,716	3.92
Intelligent home network facility installation	18,349	2.80
Miscellaneous work	6,249	0.95
Firefighting system installation	5,578	0.85
Waterproof work	4,160	0.63
Insulation work	2,464	0.38
Reinforced concrete work	1,839	0.28
Landscaping work	1,728	0.26
Network facility installation	1,101	0.17
Others	1,299	0.20
Total		656,266

나 관계를 고려하지 않고 출현 빈도만 고려하는 단어 임베딩 방법 중 하나이다. 본 연구에서는 scikit-learn 라이브러리의 CountVectorizer 모듈을 사용하여 BoW를 구현했다.

· **TF-IDF(Term Frequency-Inverse Document Frequency)**는 각 토큰이 텍스트 데이터에서 얼마나 중요한 토큰인지를 고려하는 단어 임베딩 방법 중 하나이다[9]. TF-IDF는 텍스트 데이터에서 특정 단어의 출현 빈도를 나타내는 용어 출현 빈도(TF; Term Frequency)와 특정 단어가 등장한 텍스트 데이터의 갯수를 의미하는 문서빈도(DF; Document

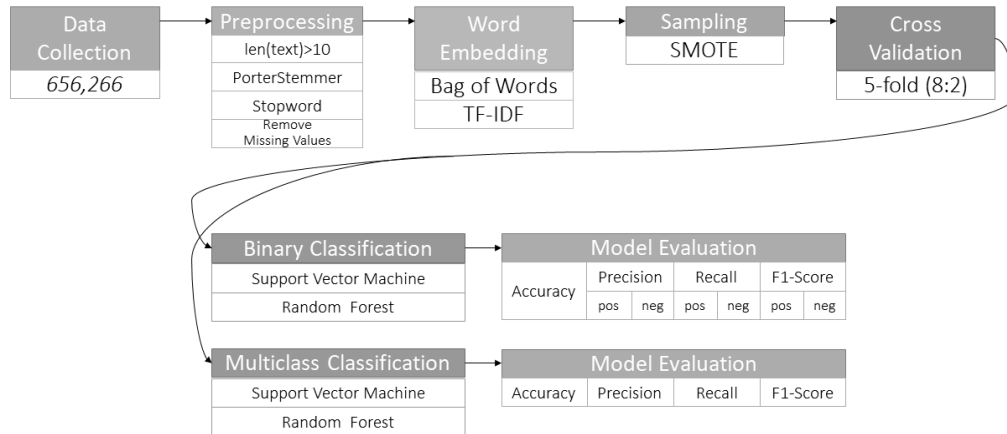


Fig. 3. Research Methodology and Procedures

Table 4. Defect Frequencies for Each Sub Categories for Finishing Work

Repair Task	Sub Category	Number of Defects	Ratio [%]	Defect Details (Example)
Finishing work	Painting work	91,979	24.65	The red stain of the entrance ceiling painting
	Tiling work	64,087	17.18	The vertical crack of the mock marble wall tile in the living room
	Paperwork	61,885	16.59	The tear in the wallpaper near the window (Worker's fault)
	Kitchen facility installation work	37,380	10.02	The gap in the joint of the sink
	Plaster finishing work	34,989	9.38	Bad condition of gypsum board behind the front door
	Indoor furniture installation work	31,517	8.45	Door imbalance of built-in cabinet (The door is not completely closed)
	Plastering work	30,060	8.06	Wall crack of internal plaster in the bathroom
	Interior masonry work	12,959	3.47	The requirement of sanding on the wall of the front-door stepping stone
	Home appliance installation work	8,262	2.21	Functional Abnormality of the wall-pad in the living room (screen freezing, malfunction)
	Total	373,118	100	

Frequency)의 역수인 역문서빈도(IDF; Inverse Document Frequency)의 곱으로 표현된다. 본 연구에서는 scikit-learn 라이브러리의 TfidfVectorizer 모듈을 이용하여 TF-IDF를 구현했다[10].

이진 분류에는 출현빈도가 높은 10,000개의 단어를, 다중 분류에는 40,000개의 단어를 선별하여 적용했다. 분류기법에 관한 상세한 설명은 3.5에서 설명할 예정이다.

3.4 샘플링(Sampling)

데이터 불균형에 따른 오류를 개선하기 위한 처리 기법인 샘플링(Sampling)은 전체 데이터 중 일부 표본을 추출하여 데이터 불균형에 따른 오류를 개선하기 위한 처리 방법이다. 대표적인 샘플링 기법으로는 무작위로 소수의 데이터를 복제하여 합성 훈련 데이터를 생성함으로써 데이터 불균형을 보완하는 기법인 오버샘플링(Oversampling)과 다수의 데이터를 간추리는 언더샘플링(Undersampling)이 있다. 이번 연구 대상인 데이터는 최대 24.65%부터 최소 2.21%까지 불균형한 분포를 보이는 라벨로 구성되어 있어 이러한 불균형 문제의 영향을 최소화하기 위해, 오버샘플링 방법 중 하나인 SMOTE 기법을 채택하여 적용했다. SMOTE란 Synthetic

Minority Oversampling Technique의 약자로 데이터 오버샘플링 기법 중의 하나이다. SMOTE는 소수 데이터마다 가장 가까운 소수 데이터 포인트를 찾아 두 포인트를 연결하는 부분에 존재하는 데이터를 임의로 샘플링하는 기법이다. 이번 연구에서는 imblearn 라이브러리의 SMOTE 모듈을 사용하여 데이터를 오버샘플링 시켰다.

3.5 분류(Classification)

분류란 결과값을 카테고리에 올바르게 매핑시키는 기술을 의미한다. 분류 결과의 유형을 기준으로 머신러닝 분류 알고리즘은 이진 분류(Binary Classification), 다중 클래스 분류(Multiclass Classification)와 다중 레이블 분류(Multi-label Classification)으로 구분된다. 본 연구에서 사용된 레이블 데이터는 하자 접수건과 9개의 상세공종과 1:1로 대응되어 있어 이진 클래스 분류와 다중 클래스 분류를 활용했다.

1) 이진 분류(Binary Classification)

이진 분류는 두 개의 클래스 중 하나로 분류하는 기술을 말한다. 세부공종 9개를 이진으로 분류하기 위해서 pandas 라이브러리의 get_dummies 모듈을 사용하여 9개의 컬럼을 생성하고 분류를 반복 수행했다[12].

본 연구에서는 분류를 위해, Support Vector Machine 과 Random Forest 모듈을 사용했다. 또한, 두 가지 분류기를 검증하기 위해 훈련 데이터와 테스트 데이터를 8:2로 나누어 5중 교차 검증을 진행했다.

Support Vector Machine은 scikit-learn 라이브러리의 LinearSVC 모듈을 이용했으며, Random Forest는 scikit-learn 라이브러리의 RandomForestClassifier모듈을 이용하여 구현했다. 그리고 scikit-learn 라이브러리의 cross_val_score 모듈을 이용하여 교차검증을 진행했다.

2) 다중 클래스 분류(Multiclass Classification)

다중 클래스 분류는 결과값을 2개 이상의 클래스에 분류하는 기술을 말한다. 다중 클래스 분류는 이진 분류와 동일한 방법으로 두 가지 분류 모델(Support Vector Machine, Random Forest)을 이용하고 이진분류와 동일한 방법으로 5중 교차검증을 진행했다.

4. 실험 결과

본 연구에서는 9개의 카테고리마다 긍정(카테고리에 해당) 과 부정(카테고리에 해당하지 않음)을 포함하는 이진 분류와 9개의 카테고리를 동시에 분류하는 다중 클래스 분류를 수행했다. TP(참 양성), FN(거짓 음성), TN(참 음성) 및 FP(거짓 양성)과 같은 지표들을 이용하여 아래 수식과 같이 정확도 (Accuracy), 정밀도(Precision), 재현율(Recall), F1점수 (F1-Score)를 계산하여 모델을 평가했다.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Table 5. Performance Metrics of Binary Classification Models

Label	Word Embedding	Classifier	Accuracy	Precision	Precision	Recall	Recall	F1-Score	F1-Score
				(positive)	(negative)	(positive)	(negative)	(positive)	(negative)
Home appliance installation	Bag of Words	Support Vector Machine	0.9771	0.9630	0.9934	0.9936	0.9618	0.978	0.9773
		Random Forest	0.9748	0.9579	0.9958	0.9960	0.9562	0.9766	0.9756
	TF-IDF	Support Vector Machine	0.9754	0.9616	0.9922	0.9925	0.9602	0.9768	0.9759
		Random Forest	0.9820	0.9688	0.9987	0.9988	0.9677	0.9836	0.9830
Paperwork	Bag of Words	Support Vector Machine	0.9544	0.9365	0.9766	0.9777	0.9337	0.9567	0.9547
		Random Forest	0.9535	0.9295	0.9863	0.9872	0.9251	0.9575	0.9547
	TF-IDF	Support Vector Machine	0.9566	0.9430	0.9731	0.9741	0.9409	0.9583	0.9567
		Random Forest	0.9666	0.9572	0.9833	0.9838	0.9559	0.9703	0.9694
Painting work	Bag of Words	Support Vector Machine	0.9448	0.9266	0.9668	0.9682	0.9235	0.9470	0.9446
		Random Forest	0.9426	0.9167	0.9792	0.9806	0.9111	0.9476	0.9439
	TF-IDF	Support Vector Machine	0.9401	0.9316	0.9516	0.9527	0.9302	0.942	0.9408
		Random Forest	0.9671	0.9724	0.9661	0.9658	0.9727	0.9691	0.9694
Plastering work	Bag of Words	Support Vector Machine	0.9416	0.9158	0.9706	0.9724	0.9105	0.9433	0.9396
		Random Forest	0.9459	0.9171	0.9853	0.9865	0.9107	0.9505	0.9466
	TF-IDF	Support Vector Machine	0.9307	0.9112	0.9580	0.9603	0.9064	0.9351	0.9315
		Random Forest	0.9652	0.9494	0.9895	0.9900	0.9473	0.9693	0.9679
Interior masonry work	Bag of Words	Support Vector Machine	0.9653	0.9476	0.9865	0.9871	0.9453	0.9670	0.9655
		Random Forest	0.9682	0.9459	0.9955	0.9958	0.9430	0.9702	0.9686
	TF-IDF	Support Vector Machine	0.9628	0.9458	0.9824	0.9832	0.9434	0.9641	0.9625
		Random Forest	0.9824	0.9693	0.9974	0.9975	0.9683	0.9832	0.9826
Plaster finishing work	Bag of Words	Support Vector Machine	0.9247	0.9010	0.9530	0.9562	0.8943	0.9278	0.9227
		Random Forest	0.9302	0.9002	0.9761	0.9783	0.8909	0.9376	0.9316
	TF-IDF	Support Vector Machine	0.9130	0.8944	0.9382	0.9410	0.8897	0.9171	0.9133
		Random Forest	0.9572	0.9531	0.9703	0.9706	0.9526	0.9618	0.9613
Indoor furniture work	Bag of Words	Support Vector Machine	0.9611	0.9427	0.9820	0.9828	0.9403	0.9623	0.9607
		Random Forest	0.9592	0.9367	0.9901	0.9906	0.9331	0.9629	0.9607
	TF-IDF	Support Vector Machine	0.9597	0.9417	0.9819	0.9827	0.9390	0.9618	0.9599
		Random Forest	0.9727	0.9578	0.9937	0.9940	0.9561	0.9756	0.9745
Kitchen facility installation work	Bag of Words	Support Vector Machine	0.9559	0.9355	0.9808	0.9819	0.9319	0.9581	0.9557
		Random Forest	0.9548	0.9304	0.9887	0.9895	0.9255	0.9590	0.9561
	TF-IDF	Support Vector Machine	0.9557	0.9388	0.9787	0.9799	0.9355	0.9589	0.9566
		Random Forest	0.9694	0.9578	0.9898	0.9903	0.9559	0.9737	0.9726
Tiling work	Bag of Words	Support Vector Machine	0.9539	0.9393	0.9729	0.9740	0.9368	0.9563	0.9545
		Random Forest	0.9535	0.9320	0.9848	0.9857	0.9278	0.9581	0.9555
	TF-IDF	Support Vector Machine	0.9548	0.9459	0.9660	0.9668	0.9446	0.9563	0.9551
		Random Forest	0.9718	0.9727	0.9737	0.9738	0.9726	0.9733	0.9731

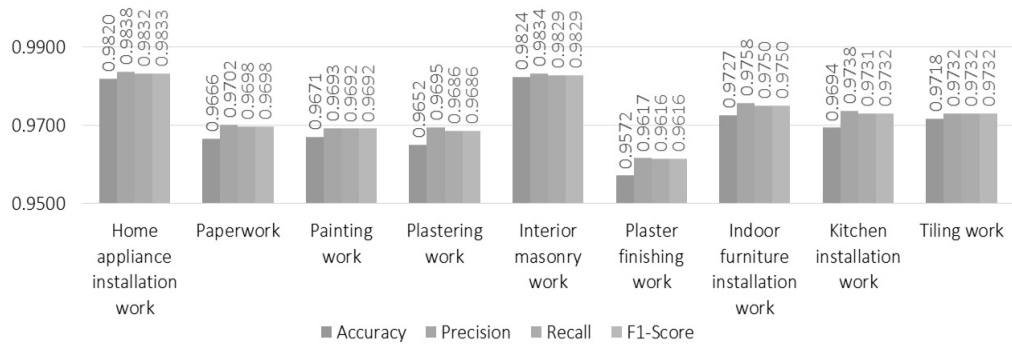


Fig. 4. Results of Binary Classification Models

Table 6. Performance Metrics of Multi-class Classification Models

Word Embedding	Classifier	Accuracy	Precision	Recall	F1-Score
Bag of Words	Support Vector Machine	0.8283	0.8430	0.8394	0.8292
	Random Forest	0.8259	0.8467	0.8387	0.8280
TF-IDF	Support Vector Machine	0.8837	0.8905	0.8894	0.8892
	Random Forest	0.9001	0.9101	0.9094	0.9092

4.1 이중 분류 결과

Table 5와 Fig. 4을 통해 이중 클래스 분류 모델의 성능을 확인할 수 있다. 그 결과, TF-IDF 단어 임베딩과 Random Forest 분류기를 사용한 모델이 가장 높은 정확도, 정밀도, 재현율 및 F1점수를 보이는 것을 확인할 수 있다.

4.2 다중 분류 결과

Table 6과 Fig. 5을 통해 다중 분류 모델의 성능을 확인할 수 있다. 그 결과, TF-IDF 단어 임베딩과 Random Forest 분류기를 사용한 모델이 가장 높은 정확도, 정밀도, 재현율 및 F1점수를 보이는 것을 확인할 수 있다.

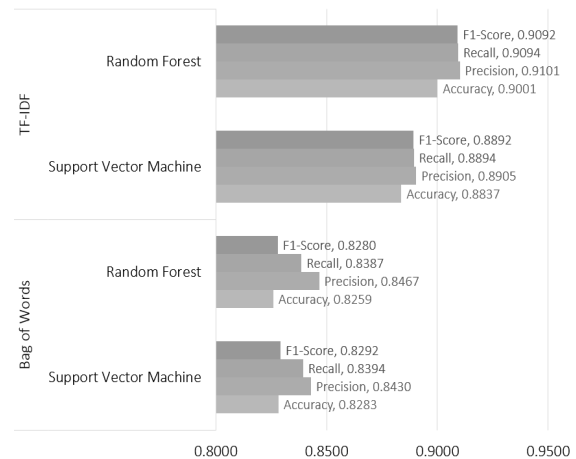


Fig. 5. Results of Multi-class Classification Models

5. 결 론

본 연구의 목표는 특정 보수 시설공사(마감공사)의 세부공종을 머신러닝 기술을 활용하여 자동 분류하는 것이다. 이러한 목표를 달성하기 위하여 대한민국 대형 아파트 건설사의 입주민 하자 접수 데이터를 수집하고 두 가지 단어 임베딩 기술(BoW, TF-IDF)과 분류 기술(Support Vector Machine, Random Forest)를 사용하여 모델을 구현했다. 그 결과, TF-IDF와 Random Forest를 이용했을 때, 이진 클래스 분류와 다중 클래스 모두 가장 높은 성능을 보였다.

상당한 정확도를 보이는 이번 연구의 머신러닝 모델을 이용하여 시공사는 경제적 비용과 인적 자원 소요를 줄이고 보다 신속하고 정확하게 보수작업 지시를 수행할 수 있다.

본 연구는 다양한 상업적이고 학문적인 의미를 내포하고 있으나 몇 가지 추가 과제와 제한사항이 남아있다. 이번 연구에서는 하나의 보수공종(마감공사)의 세부공사만을 분류하는 머신러닝 모델을 설계하고 활용하였다. 또한 연구 모델을 통

해 현장 적용 가능성을 제시하는 것에 그쳐, 추후 실제 아파트 보수현장에서 해당 모델을 적용하여 생산성을 평가하는 과제를 남겨두고 있다.

추후 연구에서는 22개의 모든 보수 시설공사(급·배수 및 위생설비공사, 창호공사, 전기 및 전력설비 공사, 냉난방·환기·공기조화 설비공사, 목공사 등)에 따른 55개의 세부공종을 분류하려는 시도를 수행해야 할 것으로 판단된다.

References

[1] Kim, Daeyeon, Housing Survey Statistical Report (2019) [Internet], <http://stat.molit.go.kr/portal/cate/statFileView.do?hRsId=327&hFormId=>

[2] D. S. Watt, "Building pathology: Principles and practice," John Wiley & Sons, 2009.

[3] Housing Construction Supply Division, Apartment Defect Dispute Mediation Committee [Internet] <http://www.adc.go.kr>

[4] Jin, Dongyeong, Apartment defect application, 62 times' explosion' in 10 years [Internet], <https://www.sedaily.com/NewsView/1Z8YOBPOY1>.

[5] The Housing Policy Division. Housing act [Internet], <https://www.law.go.kr/LSW/eng/engLsSc.do?menuId=2§ion=lawNm&query=16006>.

[6] S. H. Lee, S. H. Lee, and J. J. Kim, "Evaluating the impact of defect risks in residential buildings at the occupancy phase," *Sustainability*, Vol.10, No.12, pp.4466, 2018.

[7] B. Kim, Y. H. Ahn, and S. H. Lee, "LDA-based model for defect management in residential buildings," *Sustainability*, Vol.11, No.24, pp.7201, 2019.

[8] S. Y. Park, Y. H. Ahn, and S. H. Lee, "Analyzing the finishing works service life pattern of public housing in South Korea by probabilistic approach," *Sustainability*, Vol.10, No.12, pp.4469, 2018.

[9] T. Joachims, "A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization," *Carnegie-mellon Univ Pittsburgh Pa Dept of Computer Science*, 1996.

[10] F. Pedregosa, et al., "Scikit-learn: Machine learning in Python," *The Journal of Machine Learning Research*, Vol.12, pp.2825-2830, 2011.

[11] Glemaitre. Imbalanced-Learn [Internet], <https://github.com/scikit-learn-contrib/imbalanced-learn/tree/master/imblearn>.

[12] W. McKinney, "Data structures for statistical computing in python," *Proceedings of the 9th Python in Science Conference*, Vol.445, 2010.



김 은 혜

<https://orcid.org/0000-0003-3599-7296>
 e-mail : ekim726e@gmail.com
 2015년 인천대학교 기계공학과(학사)
 2016년 ~ 현 재 대우건설 기술연구원
 디지털건설팀 대리
 2020년 ~ 현 재 성균관대학교
 데이터사이언스융합학과 석사과정

관심분야 : 데이터과학, 인공지능활용



지 흥 근

<https://orcid.org/0000-0002-5245-5686>
 e-mail : jeehg1@skku.edu
 2020년 한국산업기술대학교 컴퓨터공학과
 (학사)
 2020년 ~ 현 재 성균관대학교
 인공지능융합학과 석사과정

관심분야 : Deep Learning for Multimedia



김 지 나

<https://orcid.org/0000-0003-1140-6113>
 e-mail : jinakim@g.skku.edu
 2019년 한성대학교 컴퓨터공학과(학사)
 2019년 ~ 현 재 성균관대학교
 인터랙션사이언스 석사과정

관심분야 : 데이터과학, 인공지능활용,
 자연어처리



박 은 일

<https://orcid.org/0000-0002-3177-3538>
 e-mail : eunilpark@skku.edu
 2010년 성균관대학교 전기전자컴퓨터공학과
 (학사)
 2012년 성균관대학교 인터랙션사이언스
 (석사)

2016년 한국과학기술원(KAIST) 기술혁신(박사)
 2015년 ~ 2017년 한국건설기술연구원(KICT) 전임연구원
 2017년 ~ 2018년 한양대학교 ERICA 조교수
 2018년 ~ 현 재 성균관대학교 인공지능융합학과 조교수
 관심분야 : 사용자경험, 데이터과학, 인공지능활용



엄 재 용

<https://orcid.org/0000-0002-7744-5572>
 e-mail : johm@kaist.ac.kr
 1977년 SUNY Buffalo Mathematics/
 Economics(학사)
 1980년 MIT Operation Research(석사)
 1986년 Northwestern University,
 Kellogg School of Management
 (박사)

1986년 ~ 1987년 Loyola University of Chicago 조교수
 1987년 ~ 1997년 AT&T Senior Member
 1997년 ~ 2001년 Lucent Technologies Senior Manager
 2001년 ~ 2008년 CapSage Consulting Director
 2009년 ~ 현 재 한국과학기술원 기술경영전문대학원 교수
 관심분야 : 기술혁신, 데이터 전략, 데이터과학, 인공지능활용

Appendix

Table 7. Examples of Defect Descriptions Written in Korean and English

Repair Work	Defect Description [KOR]	Defect Description [ENG-translated]
Finishing work - Plaster finishing work	주방 걸레받이들이 휘어져있음(원래 고정되는게 아니라 뺏다 끼웠다 하는 부분은 알고계심)/하부 물받이 프로파일 군데군데 불량임	The kitchen sinks are bent (not originally fixed, supposed to be removable and movable)/There are partial defects in the bottom of the drip tray profile.
Finishing work - Tiling work	공욕 타일 8장 제대로 부착이 안되어 퐁퐁 소리가 난다고함/부욕 타일 5장 제대로 부착 안되 퐁퐁 소리가 난다고함	8 pieces of public bathing tiles are not properly attached and it makes a plump sound / 5 pieces of bathing tiles are making a plump sound because they are not properly attached.
Finishing work - Indoor furniture installation work	맘스장 하부 측면 후면판 틈새(재접수) /후드 상부장 문짝 닫힘불량/가스렌지 상부장 문짝 수평불량	Misjang lower side rear panel gap (re-reception) / Hood upper cabinet door closed poorly / Gas stove upper cabinet door left uneven
Finishing work - Indoor furniture installation work	쌀이 밖으로 새서 보니 쌀받이 통이 없다고함(원래 그런것인 줄 알았는데 다른 세대 보니 있다함)/8/29팩스전송	When the rice leaked outside, it said that there was no rice tray (I thought it was like that, but there is another generation) / 8/29 Fax transmission
Finishing work - Indoor furniture installation work	양념장 레일 열었을때 상부쪽 깨짐/냉장고장옆 흰색 하이그로시 측면 상부쪽 하이그로시 불량-도어 600 하부카키-자재	Cracked upper side when the seasoning rail is opened/White high-gloss side of refrigerator side High-gloss upper side-door 600 Lower khaki-Material
Finishing work - Plaster finishing work	8/23오전9시이후일정예약)1.샤워부스내욕조와벽사이코킹누락/2.아일랜드식탁주위코킹누락	Schedule reservation after 8/23 9am) 1. Missing caulking between bathtub and wall in shower booth / 2. Missing caulking around kitchen table
Finishing work - Plaster finishing work	(8/2 오후)화장대 상판 거울과 맞닿은 부위 과다틈-코킹으로 처리되어있다고 불만 많음[상판을 안으로 다시 얹혀줄것요청]	(8/2 pm) Excessive gaps in the part that is in contact with the mirror on the top of the makeup table-There are many complaints that it is treated with caulking [request to re-seat the top plate inside]
Finishing work - Paperwork	현관문틀상부우측코너비트단차부위양생후샌딩후도배 예정	After curing the upper right corner bit step of the front door frame, sanding is scheduled to be painted
Finishing work - Paperwork	1/11접수건-휴오토주위 도배 불량-틈새있음//침1에어컨 주변 도배 틈새//오염다수있음 2/27(화)9~10방문	1/11 reception case-Poor wallpapering around the home auto, there are gaps// There are gaps for wallpapering around the needle 1 air conditioner//There is a lot of contamination 2/27 (Tue) 9~10 Visit
Finishing work - Indoor furniture installation work	화장대 상부 우측 거울문짝 미시공 (드레스룸) 3/10독촉-전화후 방문 -3/17독촉3/19문자보냄(295*1245)	Mirror door on the upper right side of the vanity uninstalled (dressing room) 3/10 dunning-visit after phone call-3/17 dunning 3/19 text sent (295*1245)
Finishing work - Tiling work	공용부1102동 지하2층 출입문쪽 바닥타일 2개소 손상으로 메꾸쳐리했으나 보기 흉하다며 교체요청함(타일 대리석인지 확인요청)	Common part] 2nd floor tiles on the 2nd basement floor of Building 102 were damaged due to damage, but they were asked to replace them as they were unsightly (request to confirm whether they are tiles or marble).
Finishing work - Tiling work	타일작업후공정 샤워부스내 코너선반,샤워부스코너 코킹요망 --4/6오후1시방문했으나다음일정예..	After tile work, corner shelf in shower booth, shower booth corner caulking request--4/6 I visited 1pm, but on the next schedule..
Finishing work - Plastering work	세대의부 방화문 좌측상부 크랙발생 다수-4/9일 독촉(4월은 짝수일에 보수 일정가능)	A large number of cracks in the upper left of the fire door outside the household are urged on April 9th (repair schedule available on even-numbered days in April)
Finishing work - Tiling work	공욕-양변기옆 바닥타일 구배불량/4.9일 방문건으로 완료로재접수/ 개인인테리어 줄눈세대로 인테리어줄눈 처리요청	Public bathing-Floor tile next to the toilet is poorly oriented / Re-received as a 4.9 day visit case / Request for interior joint processing by personal interior joint generation
Plumbing installation	(12/3 오전9시) 주방(공욕,부욕)-전체적으로 온수 사용시 온수쪽으로 최대돌려도 많이 뜨겁지않음(온수온도 불량)	(12/3 9:00 am) Kitchen (shared bath, fan bath)-Overall, when hot water is used, it is not very hot even if it is turned to the hot water side (poor temperature of hot water)
Plumbing installation	[모바일_접수] 세탁실-세탁기 배수시 물빠짐 불량으로 세탁실 전체에 물고임--(바닥구배, 배수구 확인요청)	[Mobile_reception] Laundry room-When the washing machine is drained, water drains into the entire laundry room due to poor drainage -- (Check the floor gradient and drain piping)
Window work	픽스창(대창)-외창 안쪽 상,하 유리코킹 (백색)약30cm 정도 누락 -직영접수-05.24 일정통화시 통화안됨	Fix window (large window)-Upper and lower glass caulking inside the outer window (white) About 30cm omission-Direct reception-May 24 No calls during certain calls
Window work	PL우창 이물질과 오염 - 창을 떼어서 작업해야함 콤팩트가 실리콘에 끼여있음, 5/19준공청소에서 PL창호로 이관	Foreign matter and contamination of PL rain window-The window must be removed. Comb is stuck in silicon, transferred from 5/19 completion cleaning to the PL window.
Electricity and power facility construction	콘센트 함몰되어 접촉 안됨/사용시 분전반 차단기 내려감/	Outlet is depressed and cannot be contacted./ When using, the switchboard breaker is lowered/
Electricity and power facility construction	전기콘센트 작동불량(썩크대 펜트리 공간 좌측 2개중 1개불량) 급하다며 빠른일정요청9/12 방문예정약속속오후3시	Malfunctioning electrical outlet (1 out of 2 on the left side of the pentree space on the sink) Request for a quick schedule 9/12 Scheduled visit 9/12 Appointment 3 o'clock
Heating, ventilation, cooling, air conditioning equipment construction	***냉매배관가스유출(에어컨설치업체에서배관연결이 처음부터잘못시공되었다며 건설사에서 처리 할 문제라고 함)	***Refrigerant pipe gas leakage (the air conditioner installation company said that the piping connection was incorrectly constructed from the beginning, and this is a problem to be dealt with by the construction company)
Heating, ventilation, cooling, air conditioning equipment construction	[모바일접수]-바람불때 주방후드 소리 심함 가스키면 가스불이 휘청거리는 정도임 11/6 10:19 팩스접수완료	[Mobile application]-When the wind blows, the sound of the kitchen hood is severe. When the gas is turned on, the gas fire is staggering. 11/6 10:19 Fax reception completed