

LLM 기반 MD&A 공시 품질 측정 방법론과 신용등급 예측에의 적용

강원모*¹, 박제민*², 함유빈*³, 오주희**

LLM-based Methodology for Measuring MD&A Disclosure Quality and Its Application to Credit Rating Prediction

Wonmo Kang*¹, Jemin Park*², Yubin Ham*³, and JooHee Oh**

요약

본 연구는 대형 언어모델(LLM)을 활용해 상장기업의 MD&A 공시품질을 항목별로 정량화하고, 신용등급 예측에의 기여도를 실증적으로 분석하였다. 금융감독원 기준에 따라 6개 항목에 대해 반복 평가하여 점수의 일관성과 해석 가능성을 확보하였다. 회귀분석 결과 '재무상태 및 영업실적'과 '분량' 항목이 신용등급과 유의미한 양(+)의 관계를 나타냈다. 투자등급/투기등급 이진 분류를 위한 머신러닝 모델 분석에서는 공시품질 변수를 포함한 XGBoost 모델의 F1 score가 9.6%, Recall이 8.6%, Precision이 10.0% 향상되어 가장 뛰어난 성능을 보였다. 본 연구는 비정형 텍스트 정보가 신용위험 예측의 정확도를 높일 수 있음을 실증적으로 확인하였으며, LLM 기반 분석이 공시문서 평가의 효율성과 정교함을 동시에 향상시킬 수 있음을 시사한다.

Abstract

This study empirically examines the role of MD&A disclosure quality in predicting corporate credit ratings using Large Language Models (LLMs). Disclosure quality of listed firms' MD&A sections was quantified across six official criteria and overall length from the Financial Supervisory Service guidelines. Repeated evaluations ensured consistent and interpretable scoring. Regression results showed that "Financial Status and Business Performance" and "Length" had significant positive effects on credit ratings. In binary classification, the XGBoost model with disclosure quality variables achieved the best performance. It improved F1 score by 9.6%, recall by 8.6%, and precision by 10.0% compared to baselines. These results confirm that unstructured textual information strengthens credit risk prediction accuracy. LLM-based evaluation further enhances efficiency and precision in disclosure assessment.

Keywords

MD&A, disclosure quality, credit rating prediction, large language model, machine learning

* 한동대학교 경영경제학부 학부과정
- ORCID¹: <https://orcid.org/0009-0001-7368-9967>
- ORCID²: <https://orcid.org/0009-0005-0534-6558>
- ORCID³: <https://orcid.org/0009-0009-8966-3890>
** 한동대학교 경영경제학부 교수(교신저자)
- ORCID: <https://orcid.org/0000-0003-0850-0256>

· Received: Oct. 29, 2025, Revised: Dec. 17, 2025, Accepted: Dec. 20, 2025
· Corresponding Author: JooHee Oh
Dept. of Management, Handong Global University, Korea
Tel.: +82-54-260-1420, Email: joohecoh@handong.edu

1. 서론

기업의 신용등급은 자본조달 비용, 외부 투자 유치, 기업 이미지 형성 등에 직결되는 핵심적인 신호로 기능한다. 특히 신용등급이 투자등급에서 투기등급으로 강등될 경우, 기업은 높은 금리로 자금을 조달해야 하며, 일부 기관투자자로부터 투자가 제한되거나 중단되기도 한다. 이는 유동성 위기와 시장 신뢰도 하락으로 이어져, 궁극적으로 기업의 장기적인 성장 가능성과 생존력에도 부정적인 영향을 미칠 수 있다[1][2]. 따라서 기업의 신용등급을 보다 정밀하게 예측하고 사전에 위험 신호를 포착하려는 시도는 실무적으로나 학문적으로나 꾸준한 관심을 받아왔다.

기존의 신용등급 예측 연구들은 주로 재무제표 기반의 정형 데이터를 중심으로 전개되었다. Altman의 Z-score 모형을 시작으로 다양한 재무 비율과 통계기법, 그리고 최근에는 머신러닝 기법을 활용한 예측모형이 개발되었지만, 이들은 주로 과거 수치를 기반으로 하며 경영진의 판단, 기업의 리스크 인식, 미래 전략 등 보다 정성적인 요소를 반영하기에는 한계가 있다. 특히 급변하는 대외 환경 속에서 기업이 스스로 인식하고 있는 위험 요소나 경영 판단의 방향성은 정형 수치로는 충분히 포착되기 어렵다.

이에 따라 최근에는 사업보고서 내 ‘이사의 경영진단 및 분석의견(MD&A)’과 같은 비정형 텍스트 정보의 활용 가능성이 주목받고 있다. MD&A는 기업 경영진이 직접 기업의 재무 상태, 영업 성과, 자금조달, 리스크 요소, 향후 계획 등을 종합적으로 기술하는 항목으로, 사업보고서 중 유일하게 경영진의 관점이 담긴 자율 공시 영역이다. 이러한 특성으로 인해 MD&A는 기업의 전략적 의도, 리스크 민감도, 정보 공개 역량 등을 드러내는 중요한 자료로 간주될 수 있으며, 정보비대칭을 완화하는 신호로도 기능할 수 있다[3].

그러나 기존의 관련 연구들은 MD&A 텍스트 전체에 대한 감성 분석에 국한되거나, 수작업 기반의 평가 방식에 의존해 왔다. 이로 인해 항목별 평가가 어렵고, 평가 기준의 일관성·재현 가능성 확보에도 한계가 존재했다. 특히 기존 방법들은 평가자의 주관이 개입될 여지가 크고, 대규모 기업을 동시에 분석하는 데 시간과 자원이 과도하게 소요되는 문제

도 있었다.

본 연구는 이러한 한계를 극복하고자, 최근 비정형 텍스트 분석 분야에서 주목받고 있는 대형 언어 모델(LLM, Large Language Model)을 활용하여 MD&A 공시품질을 자동으로 정량화하는 방법을 제안한다. 특히 금융감독원이 제시한 6개 MD&A 작성 항목에 대한 기업의 준수 정도(Degree of compliance)를 정량적으로 평가하고, 이를 기반으로 LLM 기반 공시품질 점수를 도출하였다. 평가 과정에서 판단 근거를 병렬적으로 수록하고, 반복 평가를 통해 점수의 일관성과 해석 가능성을 높였다는 점에서 기존 연구와 차별성을 지닌다.

실제로, 본 연구의 예비적 분석에 따르면 자산규모(SIZE) 평균 미만, 영업이익률(OM) 평균 미만, 또는 부채비율(TDL) 평균 초과인 기업들로 구성된 유사한 재무 건전성을 가진 집단을 대상으로 본 연구에서 제안한 LLM 기반 MD&A 공시품질 점수를 비교하였을 때, 공시품질 점수가 높은 기업 집단의 평균 신용등급이 그렇지 않은 기업 집단에 비해 뚜렷하게 높은 경향을 보였다. 그림 1은 이러한 결과를 시각적으로 보여주는데, 공시품질 평균 이상 기업들의 평균 신용등급이 A- 수준인 반면, 공시품질 평균 이하 기업들의 평균 신용등급은 BBB- 수준으로 나타나, 유사한 재무 건전성을 가진 기업들 사이에서도 공시품질에 따라 신용등급에 유의미한 차이가 존재함을 확인할 수 있다. 이는 MD&A가 단순한 서술적 문서가 아닌, 기업의 본질적 정보공개 역량과 신뢰 수준을 반영하는 요소로 작용할 수 있음을 시사한다.

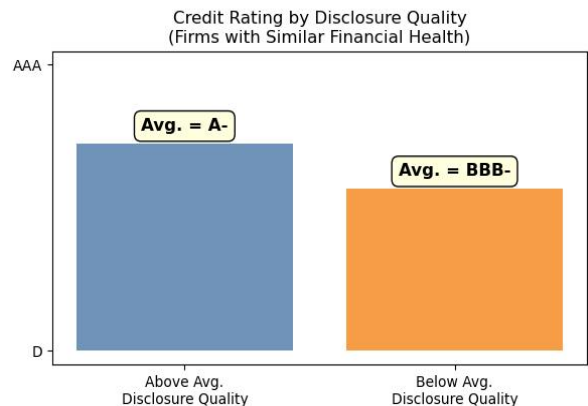


그림 1. 공시품질에 따른 신용등급 비교
Fig. 1. Comparison of credit rating by MD&A

기존 연구가 수작업 기반의 평가나 MD&A 전체에 대한 단편적인 감성분석에 그친 것과 달리, 본 연구는 LLM을 활용하여 금융감독원이 요구하는 6개의 항목들에 대한 기업들의 준수정도를 자동 분석·평가하였다. 이 과정에서 판단 근거를 병렬적으로 기록함으로써 결과의 해석 가능성을 높였으며, 반복 평가와 평균값 산출을 통해 점수의 일관성과 재현성 또한 확보하였다. 이러한 접근법은 정성적 텍스트를 체계적으로 정량화하여 기존보다 개선된 정보처리 방식을 제시한다.

본 논문은 선행연구, 연구설계, 머신러닝 예측모델, 결론의 순서로 구성되며, 정성정보 기반 신용평가의 기술적·실무적 기여 가능성을 탐구한다.

II. 선행연구

2.1 신용등급 예측 관련 선행연구

신용등급 예측에 관한 초기 연구들은 주로 재무제표 기반의 정형 수치 정보를 중심으로 이루어졌다. Altman은 다섯 가지 주요 재무비율을 조합한 Z-score 모형을 제안하여 기업의 재무 건전성과 부도 가능성을 정량적으로 평가하는 기초를 마련하였고[4], 이후 이 모델은 다양한 판별분석 기법으로 확장되며 전통적 예측모형의 기준이 되었다. 이후 김성진·안현철은 기존 통계적 방법론의 한계를 보완하기 위해 랜덤 포레스트(Random forest) 기반 머신러닝 기법을 도입해 신용등급 예측 성능을 향상시켰고, 변수 간 상호작용과 비선형성을 반영하는 모델의 유효성을 실증하였다[5]. 이근희 외는 최근 설명 가능한 인공지능(Explainable AI)을 활용한 예측모형을 제안하면서 예측력뿐 아니라 해석 가능성과 실무 적용 가능성 또한 확보하고자 하였다[6].

이러한 연구들은 신용등급 예측에서 정형 수치정보의 유용성을 실증하였지만, 경영진의 판단, 미래 전략, 리스크 인식 등은 정형 데이터로는 충분히 반영되기 어렵다. 특히 ‘이사’의 경영진단 및 분석의견(MD&A)과 같은 텍스트 기반 공시는 기업의 신용위험을 보다 정교하게 포착할 수 있는 비정형 정보로 주목된다. 이에 본 연구는 정형 재무정보 중심의

기존 접근을 확장하여, MD&A 공시 텍스트를 정량화해 예측모형에 통합하고자 한다.

2.2 공시품질 관련 선행연구

기업의 공시품질은 자본시장에서 기업의 신뢰도와 투명성, 정보비대칭 수준을 평가하는 핵심 요소로 작용한다. 이아영·김길훈은 기업의 MD&A 공시품질이 신용등급과 자본조달 비용에 유의한 영향을 미친다는 점을 실증하였으며[3], MD&A가 단순한 보고의 수단을 넘어 투자자 신뢰에 영향을 주는 질적 정보임을 강조하였다. 도용선·이다혜는 금융감독원 모범사례 준수 여부를 기준으로 MD&A의 공시구조와 표현 방식이 정보의 명확성과 가독성, 그리고 투자자 해석 용이성에 미치는 영향을 분석하였다[7]. 두 연구 모두, MD&A의 공시품질이 높을수록 기업의 정보 신뢰도가 제고되고, 시장에서 긍정적인 신호로 작용할 수 있음을 보여준다.

그러나 이러한 연구들은 대부분 수작업 기반의 평가 방식에 의존하고 있어 항목별 정보의 구조나 일관된 평가지표를 반영하는 데 한계가 있으며, 평가의 반복성과 실무 적용 가능성에도 제약이 있다. 본 연구는 이러한 한계를 극복하고자, 공시품질을 보다 구조적이고 일관된 기준으로 정량화하려는 접근을 시도한다.

2.3 LLM 기반 텍스트 정량화 관련 선행연구

최근 대형 언어모델(LLM)의 발전은 금융 분야의 비정형 텍스트 분석에 새로운 가능성을 제시하고 있다.

J. Jang and X. Wu[8]는 기업의 MD&A 전체 텍스트를 LLM에 입력하고 tone을 정량화한 후, 애널리스트의 이익예측오차와 비교하여 텍스트 감성이 실무적 유의성을 지닌다는 점을 실증하였다. 이 연구는 비영문 MD&A 문서에서도 LLM이 유의미한 정보를 추출할 수 있음을 보여준 초기 시도로, LLM의 텍스트 해석 가능성을 제시하였다.

X. Li et al.[9]은 프롬프트 기반 LLM을 활용하여 Give Me Some Credit 데이터셋의 구조화된 신용 데

이터를 자연어로 변환한 후 신용위험 분류를 수행하였다. DeepSeek-r1-distill-llama-70b 모델을 사용하여 높은 정밀도(0.91)를 달성했으며, XGBoost, Random Forest, SVM, MLP 등 전통적 머신러닝 모델과 비교 분석을 수행하였다. 특히 SHAP 분석을 통해 신용 이용률(Revolving utilization)과 연체 이력이 가장 중요한 예측 변수임을 확인하였으며, 적대적 훈련(Adversarial training)에서 LLM과 XGBoost가 유사한 견고성을 보여 LLM의 단계별 추론 과정이 해석가능성 측면에서 전통적 블랙박스 모델을 보완할 수 있음을 시사하였다.

Y. Lee et al.[10]은 기업 신용평가에서 비재무적 정성 지표를 기반으로 구조화된 추론을 생성하는 토론 기반 다중 에이전트 시스템(KPD-MADS)을 개발하였다. Karl Popper의 비판적 대화 프레임워크에 기반한 10단계 상호작용 프로토콜을 통해, 단일 에이전트 시스템 대비 설명적 적절성(중양값 4.0 vs 3.0)과 실용적 적용 가능성(4.0 vs. 3.0)에서 우수한 평가를 받았으며, 수동 평가(1920초) 대비 약 95%의 시간 단축(91.97초)을 달성하였다.

U. O. Ogbuonyalu et al.[11]은 LLM이 은행 리스크 모델에서 거래 내역, 고객 커뮤니케이션, 소셜미디어 감정, 금융 뉴스 등 다양한 비정형 데이터를 분석하여 차용자 행동과 신용도에 대한 깊은 통찰을 제공할 수 있다는 포괄적 리뷰를 제시하였다. JPMorgan Chase의 계약서 분석 자동화(36만 시간에서 수 초로 단축), Goldman Sachs의 실적발표 감정 분석, Capital One의 행동 기반 신용평가 등 실무 사례를 통해, LLM이 리스크 예측 정확도를 개선하고 금융소외계층에 대한 포용적 평가를 가능하게 함을 실증하였다.

이처럼 LLM을 활용한 금융 텍스트 분석 연구가 활발히 진행되고 있으나, 대부분의 연구는 감성 분석, 리스크 예측, 정보 추출에 집중되어 있으며, 텍스트로부터 시장 관련 신호를 도출하거나 이미 구조화된 데이터의 분류 성능을 개선하는 데 초점을 맞추고 있다. 그러나 MD&A의 '공시품질' 자체, 즉 공시 문서가 규제 기준을 얼마나 충실히 준수하고 항목별 정보가 얼마나 완결적으로 작성되었는지를 LLM을 활용하여 체계적으로 정량화한 연구는 아직까지 시도된 바 없다. 본 연구는 이러한 연구 공백

을 메우고자, 금융감독원이 제시한 6개 MD&A 작성 항목에 대한 기업의 준수 정도를 항목별로 구조화하여 정량화하는 방법을 제안한다. 이를 통해 정형 재무정보 중심의 신용등급 예측모형[4]-[6]에 정성적 공시품질 정보를 체계적으로 통합하여, 그림 1에서 제시한 바와 같이 유사한 재무 건전성을 가진 기업들 사이에서도 공시품질에 따른 신용등급 차이를 설명할 수 있는 정교한 분석 프레임워크를 제시하고자 한다.

III. 연구설계

본 장에서는 본 연구의 핵심인 MD&A 텍스트 공시품질과 기업 신용등급 간의 관계를 실증적으로 분석하기 위한 연구 설계 및 방법론을 상세히 기술한다. 데이터수집 과정, 변수 설정, 그리고 분석 모형 구축에 이르는 전반적인 절차를 명확하게 제시하여 연구의 투명성과 재현 가능성을 확보하고자 한다. 특히, 대규모 언어 모델(LLM)을 활용한 MD&A 공시품질 정량화 과정은 본 연구의 차별화된 접근 방식으로서, 그 신뢰성과 타당성을 확보하기 위한 구체적인 방법론을 심도 있게 다룬다.

3.1 데이터수집

본 연구는 281개 상장기업(금융업 제외)을 대상으로, 2023년 회계연도 기준 MD&A 텍스트 및 재무 변수를 독립변수로, 그리고 이듬해인 2024년의 신용등급을 종속변수로 설정하여 분석하였다. 이러한 시차를 둔 변수 설정은 MD&A 공시 정보가 기업의 미래 신용도에 미치는 영향을 보다 명확하게 파악하기 위함이다. 데이터는 FnGuide와 금융감독원 DART에서 수집하였다[12][13]. 신용등급은 AAA부터 D까지 21단계를 0~21로 변환하였으며, 이러한 수치화는 신용등급 간의 상대적 차이를 정량적으로 분석하고 회귀분석에 용이하게 활용하기 위한 조치이다. 재무 변수는 기업규모(SIZE), 총부채비율(TDL), 영업이익률(OM), 시장가치 대비 장부가치비율(MB), 배당 여부(DIVP)를 사용하였다(G. H. Lee et al.[6]). 구체적으로, 기업규모(SIZE)는 자연로그를 취한 시가총액으로, 기업의 전반적인 규모와 시장

영향력을 대변한다. 총부채비율(TDL)은 총부채를 총자산으로 나눈 값으로, 기업의 재무 레버리지 및 부채 상환 능력을 나타낸다. 영업이익률(OM)은 영업이익을 매출액으로 나눈 값으로, 기업의 핵심 영업 활동을 통한 수익성을 평가한다. 시장가치 대비 장부가치 비율(MB)은 시가총액을 순장부가치로 나눈 값으로, 기업의 성장 잠재력과 시장의 기대치를 반영한다. 마지막으로 배당 여부(DIVP)는 배당 지급 유무를 나타내는 더미 변수로, 기업의 재무적 안정성과 주주 환원 정책을 간접적으로 시사한다.

MD&A 텍스트는 Python 라이브러리를 활용해 DART에서 크롤링한 후 전처리하였으며, 공시품질 평가는 LLM인 GEMINI 2.0 Flash Thinking 모델을 활용하여 수행하였다. 그림 2는 본 연구의 MD&A 텍스트 수집 및 공시품질 정량화 과정의 전체 워크플로우를 나타낸다. 이는 기존의 수동 평가 방식이 지닌 시간과 인적 자원 소모의 한계를 극복하고, 대량의 텍스트 데이터를 보다 일관적이고 객관적인 기준으로 평가하기 위함이다. 또한, 일반적인 GEMINI 2.0 Flash 모델보다 추론 성능이 강화된 Flash Thinking 모델을 활용하는 것이 평가의 일관성과 신뢰도 측면에서 더 우수한 결과를 보였다. 평가 기준은 금융감독원의 '기업공시 길라잡이'에 수록된 '이사의 경영진단 및 분석의견' 항목 지침을 참고하였다. 이 지침은 MD&A 공시의 목적, 포함되어야 할 내용, 그리고 작성 원칙 등을 명시하고 있어, 공시품질을 평가하기 위한 신뢰성 있는 근거를 제공한다. 이에 따라 6개 항목'에 대한 프롬프트를 구

성하고, 자동 평가를 통해 각 항목별로 0점에서 6점까지 점수를 부여하였다. 그림 3은 LLM 평가에 사용된 프롬프트의 예시를 보여준다. 또한, MD&A의 총 낱말 수에 자연로그를 취해 공시의 양적 측정치인 'Length' 변수를 생성하였다. 이는 선행연구에서도 자발적 공시수준의 대용치로 활용된 바 있다[3].

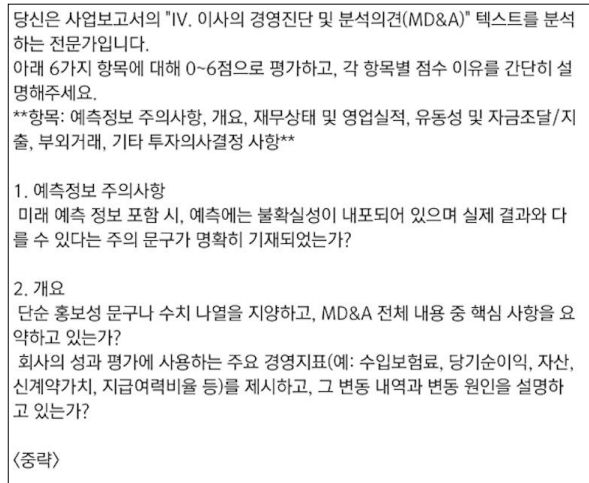


그림 3. 프롬프트 예시
Fig. 3. Example of the prompt

LLM 평가의 일관성과 신뢰도를 확보하기 위해, 모델의 하이퍼파라미터인 temperature 값은 0.3으로 설정하였다. Temperature는 0에서 2 사이의 값을 가지며, 값이 클수록 더 창의적이고 다양성이 높은 응답을 생성한다. 하지만 본 연구에서는 평가의 일관성과 재현 가능성을 중시하여 비교적 낮은 값으로 설정하였다.

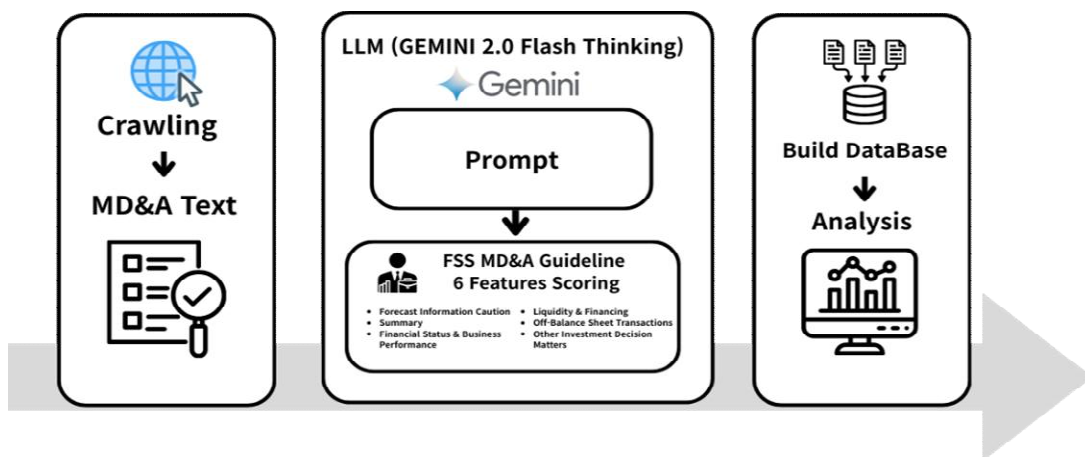


그림 2. MD&A 텍스트 수집 및 공시품질 정량화 과정
Fig. 2. Workflow of the data collection process

또한, 각 MD&A 항목에 대해 20회 이상의 반복 평가를 수행하여 평균 점수를 최종 점수로 사용하였다. 이러한 반복 평가는 LLM의 확률적 특성으로 인한 평가 편차를 최소화하고, 안정적인 평가 결과를 도출하기 위한 통계적 방법이다. 실제로, 반복 평가 결과 항목별 점수의 표준편차가 평균 8% 이상 감소하여 평가의 안정성을 통계적으로 확인하였다(그림 4). 또한, LLM이 각 항목을 평가할 때 제시한 판단 근거를 병렬적으로 생성하도록 설계하였다[14]. 이는 LLM의 평가 결과를 단순히 수용하는 것을 넘어, 평가의 배경과 논리를 이해하고 필요시 검증할 수 있는 기반을 제공하여, 공시 텍스트와의 정합성을 강화하고 LLM 평가의 해석 가능성과 투명성을 제고하는 데 기여한다.

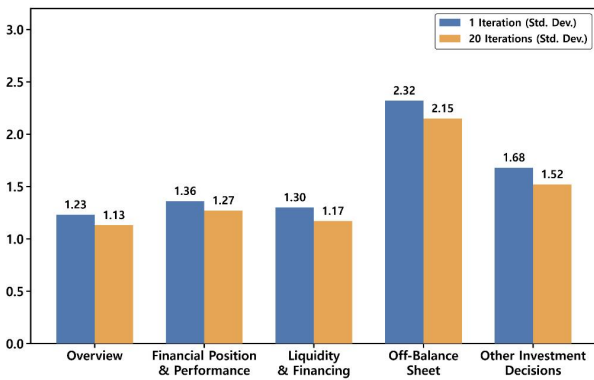


그림 4. 반복 횟수별 표준편차 비교
Fig. 4. Standard deviation by number of repetitions

3.2 회귀분석

본 연구는 MD&A 공시품질 변수들이 기업의 신용등급에 통계적으로 유의미한 영향을 미치는지를 실증적으로 분석하기 위해 다중 회귀분석 모형을 구축하였다. 회귀분석은 다양한 독립변수들이 종속 변수에 미치는 개별적인 영향을 통계적으로 추정하고, 변수들 간의 관계를 수량화하는 데 효과적인 방법이다. 본 연구에서는 MD&A 공시품질과 관련된 정성적 항목들(LLM 평가 점수)을 주요 독립변수로 설정하고, 수치형으로 변환된 신용등급을 종속변수로 설정하였다. 회귀식은 다음과 같다.

$$\begin{aligned}
 CreditRating = & \beta_0 + \beta_1 \cdot SIZE + \beta_2 \cdot TDL + \beta_3 \cdot OM \\
 & + \beta_4 \cdot MB + \beta_5 \cdot DIVP + \beta_6 \cdot Length \\
 & + \beta_7 \cdot ForecastNote + \beta_8 \cdot Summary \\
 & + \beta_9 \cdot FinPerf + \beta_{10} \cdot Liquidity \\
 & + \beta_{11} \cdot OffBS + \beta_{12} \cdot OtherInfo + \epsilon \quad (31)
 \end{aligned}$$

회귀분석 결과(표 1 참조), MD&A 공시품질 항목 중 '재무상태 및 영업실적'($\beta=0.54, p=0.029$)과 MD&A의 양적 측정치인 '분량'($\beta=0.55, p=0.023$)이 기업 신용등급과 통계적으로 유의미한 양(+)의 관계를 나타냈다. 이는 기업이 MD&A를 통해 손익, 자산 구성, 산업 내 위험 요소 등에 대해 얼마나 체계적이고 상세한 설명을 제공하는지, 그리고 충분한 정보를 제공하는지가 외부 신용평가에서 긍정적인 신호로 작용함을 강력히 시사한다. 즉, 재무 상태와 영업 실적에 대한 충실하고 깊이 있는 설명은 기업의 투명성과 재무 건전성에 대한 신뢰를 높여 신용등급 상향에 기여할 수 있으며, 공시 분량이 많을수록 정보 비대칭이 감소하여 신용평가기관의 판단에 긍정적인 영향을 미칠 수 있음을 의미한다.

표 1. 신용등급 회귀분석 결과
Table 1. Regression analysis results of credit rating

Variable description	Variable description	coefficient	p-value
Inercept	Intercept	14.59***	0.000
SIZE	Market capitalization (log)	2.06***	0.000
TDL	Total debt/total assets	-0.54***	0.007
OM	Operating income/revenue	0.44**	0.030
MB	Market cap/net book value	-0.23	0.119
DIVP	Dividend payout status	1.33***	0.000
Length	ln(MD&A word count)	0.55**	0.023
Forecast note	Forecast information caution	0.12	0.403
Summary	Summary	0.18	0.305
FinPerf	FinancialStatus & business performance	0.54**	0.029
Liquidty	Liquidity & Financing	-0.28	0.180
OffBS	Off-balance sheet transactions	0.29*	0.086
OtherInfo	Other investment decision matters	-0.30	0.117
Number of observations		281	
Adjusted R		0.710	
F-statistic		58.01	
P-value		p<0.001	

이러한 회귀분석 결과를 바탕으로, 본 연구는 분석을 확장하여 투기등급(0)과 투자등급(1)을 구분하는 이진 분류모델로의 적용 가능성을 탐색하였다. 신용등급의 하락, 특히 투기등급으로의 전환은 기업에게 자금조달 비용 상승, 대외 신인도 하락, 심지어 생존 가능성 저하와 같은 실질적이고 심각한 리스크를 초래할 수 있다[2][15]. 따라서 기업의 신용위험을 사전에 예측하고 관리하는 것은 기업 경영 및 투자 결정에 있어 매우 중요한 과제이다.

그림 5는 투기등급 기업과 투자등급 기업 간의 MD&A 공시품질 점수 차이를 시각적으로 명확하게 보여주는 자료이다. 이 그림은 각 등급 간 평균 공시품질 점수 차이가 통계적으로 유의미하게 나타남을 보여주며, 이는 MD&A 공시품질이 기업의 신용위험을 반영하는 유의미한 신호임을 재확인시켜 준다. 이러한 결과를 바탕으로, 본 연구는 재무변수와 회귀분석에서 유의미하게 나타난 MD&A 공시품질 항목들을 결합하여, 기업의 신용위험(투기등급 전환 가능성)을 예측하는 분류 예측 모델을 설계하였다. 이 모델은 재무 정보와 비재무적 정보인 MD&A 공시품질을 통합하여 기업의 신용도를 보다 포괄적으로 평가하고 예측하는 새로운 방법론을 제시한다.

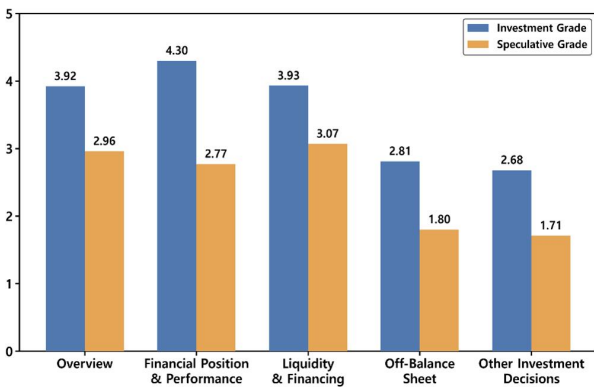


그림 5. MD&A 평균 점수 비교: 투자등급과 투기등급
Fig. 5. Comparison of MD&A scores by credit grade

IV. 머신러닝 예측 모델

본 연구에서는 기업의 신용등급을 투자등급(1)과 투기등급(0)으로 단순화하여, 이를 이진 분류하는 머신러닝 기반 예측모델을 설계하였다. 전체 표본 281개 기업 중 투자등급 기업은 223개(79.4%), 투기

등급 기업은 58개(20.6%)로 구성되어 클래스 불균형이 존재한다. 데이터의 개수가 부족한 점을 보완하고 모델 성능의 신뢰성을 확보하기 위해 k-fold 교차검증(k-fold cross-validation) 기법을 적용하였다. 이 방법은 전체 데이터를 일정한 크기의 부분집합으로 나누는 뒤, 각 subset을 한 번씩 검증용으로 사용하고 나머지를 학습에 사용하는 과정을 반복함으로써 모델을 보다 일반화 가능하게 평가하는 절차이다. 본 연구에서는 k=5로 설정하여 총 5회의 반복 학습과 검증을 수행하였으며, 각 폴드의 훈련/검증 세트에서도 투자등급 약 79%, 투기등급 약 21%의 비율이 유지되었다. 분석에는 두 가지 예측모형이 활용되었다. 첫 번째는 재무정보만을 포함한 기준모형(Model 1), 두 번째는 여기에 LLM 기반 공시품질 점수를 결합한 확장모형(Model 2)이다(표 2 참조).

표 2. 머신러닝 모델 결과

Table 2. Results of machine learning models

Model	Metric	Model1 (Financial only)	Model2 (Financial+ MD&A)	Δ(Max)
Logistic	Recall	0.64	0.70	9.4%
	Precision	0.82	0.82	
	F1	0.71	0.76	
SVM	Recall	0.58	0.69	19.0%
	Precision	0.73	0.78	
	F1	0.64	0.73	
RF	Recall	0.65	0.72	10.8%
	Precision	0.83	0.84	
	F1	0.71	0.76	
XGB	Recall	0.70	0.76	10.0%
	Precision	0.80	0.88	
	F1	0.73	0.80	

분석 결과, 공시품질 정보를 포함한 모든 확장모형은 기준모형 대비 전반적인 분류 성능 향상을 나타냈다. 특히 투기등급 기업의 탐지 성능(Recall)의 유의미한 개선이 관찰되었는데, 이는 신용등급 하락 위험이 높은 기업을 조기에 식별하는 데 중요한 함의를 갖는다. 이와 함께, Precision, F1 Score 등 다른 주요 성능 지표들 역시 모든 확장모형에서 고르게 향상된 것으로 나타났다. 특히 XGBoost 모델은 공시품질 정보를 포함한 이후 가장 뚜렷한 성능 개선을 보였으며, Recall 8.6%, Precision 10.0%, F1

Score 9.6%, Accuracy 3.3%의 상승폭을 기록하였다. 이는 단순한 탐지 범위의 확대를 넘어, 모델이 보다 정밀하고 신뢰도 높은 방식으로 신용위험을 분류할 수 있게 되었음을 시사한다.

이러한 결과는 기업의 정성적 정보, 특히 경영진이 서술한 재무 및 사업 현황에 대한 설명 수준과 공시 분량이, 외부 평가자에게 신뢰도 높은 신호로 작용함을 시사한다. 다시 말해, MD&A의 충실한 공시는 기업의 실질적 위험과 전망을 보다 명확하게 전달하며, 이는 머신러닝 분류에서도 신용위험을 판별하는 데 기여할 수 있음을 실증적으로 보여준다. 특히 본 연구는 공시품질 점수가 단지 보조 설명변수를 넘어, 실질적인 분류 성능 개선으로 이어지는 유효한 정보임을 확인하였다는 점에서 기존 재무 중심 예측모형의 한계를 보완하는 대안적 접근으로서 의미를 가진다.

V. 결 론

본 연구는 대형 언어모델(LLM)을 활용하여 기업의 MD&A 공시품질을 항목별로 정량화하고, 이를 신용등급 예측모형에 통합함으로써 비정형 텍스트 정보가 신용평가에 유의미하게 활용될 수 있음을 실증적으로 검증하였다. 전통적인 신용등급 예측이 주로 재무제표 기반의 정형 수치에 의존해온 데 반해, 본 연구는 공시문서 내 질적 정보, 특히 경영진의 전략 판단과 미래전망이 담긴 MD&A 텍스트가 신용위험 판단에 실질적인 영향을 미친다는 점에 주목하였다.

첫째, 실증 결과 MD&A 공시품질 항목 중 ‘재무상태 및 영업실적’과 문서 ‘분량’은 신용등급과 통계적으로 유의한 양의 관계를 나타냈다. 이는 기업이 자신의 재무성과와 운영 성과에 대해 투명하고 구체적으로 서술할수록 시장과 평가기관으로부터 더 높은 신뢰를 얻을 수 있음을 보여준다. 특히 ‘분량’이라는 단순 지표조차 정보 충실도의 대리변수로 작용할 수 있음을 확인함으로써, 향후 공시문서 평가에 있어 새로운 정량지표의 가능성을 제시하였다.

둘째, 재무정보에 공시품질 변수를 결합한 머신

러닝 기반 예측모형은 기존 재무정보 기반 모델에 비해 전반적인 예측 성능이 향상되었으며, 특히 투기등급 기업에 대한 탐지 능력이 뚜렷이 개선되었다. 이는 공시의 질적 수준이 높은 기업일수록 장기적인 재무 안정성과 신용위험 관리 역량도 우수하다는 가설을 뒷받침하며, LLM 기반 텍스트 분석이 실질적인 리스크 조기 경고 체계로 활용될 수 있음을 시사한다. XGBoost 모델의 경우 F1 score 9.6%, Recall 8.6%, Precision 10.0%, Accuracy 3.3%의 성능 향상을 기록하며 가장 두드러진 개선 효과를 보였다.

셋째, 본 연구는 기존 선행연구들이 가지는 몇 가지 핵심 한계를 실증적으로 보완하였다. 기존의 공시품질 연구들은 수작업 평가에 의존하거나 문서 전체를 단일 점수화하는 방식에 머무르며 항목별 정보 반영력, 일관성, 재현성 측면에서 제약이 있었다. 반면, 본 연구는 금융감독원이 제시한 MD&A 항목별 구조를 바탕으로 프롬프트를 설계하고, 항목별 반복 평가를 통해 통계적 신뢰성과 정밀도를 확보하였다. 이는 공시 텍스트의 구조적 정보를 체계적으로 정량화한 첫 시도 중 하나로, 학문적으로도 새로운 분석 프레임워크를 제시한다.

이러한 연구 결과는 학문적·실무적 측면에서 다음과 같은 의의를 가진다. 학문적으로는 신용위험 평가에 있어 비정형 데이터의 활용 가능성을 실증하였으며, LLM 기반 텍스트 분석이 정성적 정보를 정량화하는 강력한 도구가 될 수 있음을 보여주었다. 실무적으로는 기업의 공시품질 개선이 단지 규제 준수 차원을 넘어, 실질적인 신용등급 관리 전략이 될 수 있다는 점을 시사하며, 투자자·평가기관·감독기관 등 정보 이용자들에게 공시문서 평가 자동화의 가능성과 방향성을 제시한다.

물론 본 연구는 단일 회계연도 데이터와 일부 산업 편향이라는 제한점이 존재하며, 텍스트 내 문맥적 맥락을 완전히 반영하지 못한 한계도 있다. 향후 연구에서는 다년간의 시계열 데이터를 활용하여 텍스트 점수의 시간적 안정성을 검토하고, 산업별 맞춤형 분석을 통해 일반화 가능성을 높일 수 있을 것이다. 또한 다양한 LLM 구조와 평가 방식에 대한 비교 연구, 프롬프트 최적화 기법 도입 등을 통

해 평가 신뢰성과 실무 적용력을 더욱 강화할 수 있을 것이다.

궁극적으로 본 연구는 정형 재무정보와 정성 텍스트 정보를 통합한 신용등급 예측의 새로운 길을 제시하며, 공시 텍스트의 실질적 가치를 발굴하고 이를 계량화하는 첫 실험적 시도로서, 향후 LLM 기반 기업분석 및 위험관리 분야의 확장적 연구에 중요한 초석이 될 수 있을 것이다.

References

- [1] M. T. Kim and J. B. Wi, "The impact of auditor size and audit opinion on corporate credit ratings", *Journal of Accounting and Auditing Research*, Vol. 35, No. 5, pp. 45-74, Jun. 2006.
- [2] T. K. Kim and J. S. Shin, "Credit Ratings and Corporate Investment Decisions", *Korean Journal of Securities Studies*, Vol. 46, No. 4, pp. 785-807, Sep. 2017. <http://dx.doi.org/10.26845/KJFS.2017.09.46.4.785>.
- [3] A. Y. Lee and G. H. Kim, "The Disclosure Quality of Management Discussion and Analysis and the Cost of Debt", *Journal of Business Education Research*, Vol. 29, No. 1, pp. 407-430, 2014. <https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001856990>.
- [4] E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, Vol. 23, No. 4, pp. 589-609, Aug. 1968. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>.
- [5] S. J. Kim and H. C. Ahn, "Application of Random Forest for Corporate Credit Rating Prediction", *Industry and Innovation Research*, Vol. 32, No. 1, pp. 187-211, 2016. <https://doi.org/10.22793/indinn.2016.32.1.006>
- [6] G. H. Lee, K. M. Kim, and H. J. Lim, "Study on Predicting Corporate Credit Ratings Using CART", *Journal of the Korean Data and Information Science Society*, Vol. 35, No. 5, pp. 585-596, Jun. 2024. <https://doi.org/10.7465/jkdi.2024.35.5.585>.
- [7] Y. S. Do and D. H. Lee, "The Impact of Referencing MD&A Best Practices on Information Asymmetry and Readability", *Korean Accounting Journal*, Vol. 22, No. 1, pp. 1-22, Feb. 2024. <https://doi.org/10.24056/KAJ.2023.11.002artiId=ART001141373>.
- [8] J. Jang and X. Wu, "Non-English Textual Analysis with Large Language Models: Analysts' Use of MD&A Sentiment in Earnings Forecasting", *SSRN Electronic Journal*, pp. 1-45, Jan. 2024. <https://doi.org/10.2139/ssrn.5123384>.
- [9] X. Li, Y. Zhang, and H. Chen, "Explore the use of prompt-based LLM for credit risk classification", *Journal of Computer and Communications*, Vol. 13, No. 6, pp. 33-46, Jun. 2025. <https://doi.org/10.4236/jcc.2025.136003>.
- [10] Y. Lee, M. Kim, and H. Choi, "Structured debate improves corporate credit reasoning in financial AI", *arXiv preprint, arXiv:2510.17108*, pp. 1-12, Oct. 2025. <https://doi.org/10.48550/arXiv.2510.17108>.
- [11] U. O. Ogbuonyalu, K. Abiodun, and S. Dzemefe, "Beyond the credit score: The untapped power of LLMs in banking risk models", *Finance & Accounting Research Journal*, Vol. 7, No. 4, pp. 183-204, May 2025. <https://doi.org/10.51594/farj.v7i4.1905>.
- [12] FnGuide, "FnGuide Database", <https://www.fnguide.com>. [accessed: Jun. 01, 2025]
- [13] Financial Supervisory Service, "Data Analysis, Retrieval and Transfer System (DART)", <https://dart.fss.or.kr>. [accessed: Jun. 01, 2025]
- [14] S. Y. Jung, "Predicting Corporate Credit Rating Using RAG-based LLM: Focusing on Corporate Credit Rating Explanation and Outlook", M.S. Thesis, Sogang University, Seoul, Korea, 2024. <http://www.dcollection.net/handler/sogang/000000078976>.
- [15] D. W. Kim and J. Ahn, "Issue of Junk Bonds in Flight to Quality", *Korean Journal of Financial Management*, Vol. 35, No. 1, pp. 1-25, 2018. <https://doi.org/10.22510/kjofm.2018.35.1.001>.

저자소개

강 원 모 (Wonmo Kang)



2022년 3월 ~ 현재 :
한동대학교 경영경제학 학부과정
(경영학&AI융합학 전공)
관심분야 : 재무분석, 금융데이터
분석, 인공지능

박 제 민 (Jemin Park)



2020년 3월 ~ 현재 :
한동대학교 경영경제학 학부과정
(경영학&AI융합학 전공)
관심분야 : 설명가능한
인공지능(XAI), 금융데이터 분석,
대형 언어 모델(LLM), 머신
러닝

함 유 빈 (Yubin Ham)



2019년 3월 ~ 현재 :
한동대학교 경영경제학 학부과정
(경영학&AI융합학 전공)
관심분야 : 설명가능한
인공지능(XAI), 대형 언어
모델(LLM), 머신러닝,
비즈니스 애널리틱스

오 주 희 (Joohee Oh)



2004년 3월 :
서울대학교 경제학부(석사)
2011년 8월 : University of
Southern California(경영학박사)
2021년 3월 ~ 현재 : 한동대학교
부교수
관심분야 : 디지털 경제,

정보시스템, 비즈니스 애널리틱스, 금융데이터마이닝