창업 기업의 지속 가능성 강화를 위한 데이터 분석 방법론 제안

김지예 · 김원빈*

한국성서대학교

Data Analysis Methodology to Improve Startup Sustainability

Jiye Kim · Wonvin Kim*

Korean Bible University

E-mail : jiye3180@bible.ac.kr / wkim@bible.ac.kr

요 약

비즈니스 환경의 불안정성으로 국내 창업 기업의 생존율이 저조한 상황이며, 이를 극복하기 위하여 효율적인 비즈니스 모델의 필요성이 증가되고 있다. 이에, 본 논문은 창업 기업의 지속 가능성을 높이기 위한 데이터 분석 방법론을 제안한다. 첫 번째로, 창업 기업 및 경쟁사의 공시자료를 수집 후 기업의 성장, 수익, 안정, 현금흐름 등을 수치화하고 강점과 약점 요인을 도출한다. 두 번째로, RSI(relative strength index) 지표를 이용하여 과매수 및 과매도 시점들을 파악 후 해당 기간의 뉴스 기사들을 수집하고, FastText 모델을 기반으로 기회와 위협 요인들로 분류한 어휘 사전을 생성한다. 생성된 어휘 사전과 로지스틱 회귀 모델을 기반으로 해당 기업의 기회와 위협 요인의 분석을 수행한다. 이를 통해, 창업 기업의 비즈니스 전략 수립에 기여한다.

ABSTRACT

Due to the instability of the business environment, the survival rate of domestic startups is low. To overcome this, the need for efficient business models is increasing. This paper proposes a data analysis methodology to increase the sustainability of startups. Fisrst, the proposed methodology quantifis a company's growth, profitability, stability, cash flow, etc. and identifies strength and weakness factors, after collecting disclosure data from startup and competitors. Second, after identifying overbought and oversold points using the Relative Strength Index (RSI) indicator, news articles from the period are collected. Then, based on the FastText model, a vocabulary dictionary is generated, divided into opportunity and threat factors. And then, an analysis of the opportunity and threat factors of the startup is carried out based on the vocabulary dictionary generated and the logistic regression model. In conclusion, this paper confirms the theoretical basis and practical applicability for increasing the sustainability of resource-limited startups.

키워드

Startup, Venture, SWOT, FastText, TF-IDF

Ⅰ. 서 론

중소벤처기업부(2021) 통계에 따르면 국내 창업 기업의 1년 생존율은 64.8%에 불과하고, 5년 생존율은 33.8% 수준에 그치고 있다. 특히, 1년 이상 살아남은 기업조차도 5년간 지속률은 52.27%로 매우 낮다. 이러한 창업 실패의 주된 원인은 자금 확보의 어려움, 실패에 대한 두려움, 지식·능력·경험의 부족 등이다.

본 연구에서는 창업 기업의 지속 가능성을 높이기 위해 비즈니스 모델의 SWOT 분석을 수행하는 새로운 방법론을 제안한다. SWOT 분석은 기업의 내·외부 환경 요인을 객관적으로 분석하는 도구로, 기업의 현재 위치 진단과 향후 전략 수립에 필요한 정보 확보에 도움이 된다[1].

본 연구에서는 비교 대상 기업들의 공시 자료[2]

* corresponding author

를 참고하여 재무 상태 및 사업 현황을 비교함으로써, 내부 요인인 강점(strength)과 약점(weakness)을 도출한다. 그리고 비교 대상 기업들의 주식 과매수 및 과매도 시점들을 상대강도지수(RSI: relative strength index) 지표를 이용하여 파악하고, 해당 기간의 뉴스 기사를 수집하여 외부 요인인 기회(opportunity)와 위협(threat)을 도출한다.

II. 관련 연구

1) FastText

FastText는 Facebook에서 개발한 예측 기반 단어 임베딩 모델로 각 단어를 문자 단위의 n-gram으로 표현하고 n-gram 벡터를 학습한다[3]. 이 모델은 Word2Vec 기술이 단어의 내부 구조를 고려하지 않는 것과 달리, 단어의 형태학적 특성을 활용하여 신조어나 오탈자가 포함된 단어에 대해서도 유사도를 효과적으로 계산할 수 있다는 장점이 있다. 이는 어휘의 다양성이 높은 언어나 구조가 복잡한 문서를 다룰 때 유용하다. 또한, FastText는 문자 단위의 정보 활용으로 인해 딥러닝 기반 자연어 처리 모델보다 속도가 상대적으로 빠르다는 장점이 있다.

2) TF-IDF

TF-IDF(term frequency-inverse document frequency)는 문서 내 특정 단어의 중요도를 평가하는 데 사용되는 가중치이다[4]. 'TF'는 문서 내에서 단어가 등장하는 빈도를 나타내며, 'IDF'는 전체 문서 집합에서 해당 단어가 얼마나 드물게 나타나는지를 측정한다. 'TF'와 'IDF'의 곱은 단어의 문서 내 중요도를 수치화한다. TF-IDF는 문서의 주요 주제를 식별하거나 문서 간의 유사성을 평가하는 데 주로 활용된다.

3) 코사인 유사도

코사인 유사도(cosine similarity)는 두 벡터 간 코사인을 이용하여 유사성을 측정하는 표준적인 방법이다. 이 방법은 벡터의 길이보다는 방향에 중점을 두어 두 벡터의 유사도를 0에서 1 사이의 값으로 표현한다. 유사도 값이 1에 가까울수록 두 벡터는 유사한 방향을 가지는 반면 0에 가까울수록 방향이 상당히 다르다는 것을 의미한다. 이는 고차원 데이터 분석, 텍스트 데이터의 유사도 측정, 문서 분류 작업 등에서 활용된다.

III. 제안 방법론

1) 강점, 약점 분석

공시 자료를 수집하여 데이터를 전처리한다. 전처리된 데이터로 시장 내에서 경쟁사 대비 자사의 성장성, 수익성, 안정성, 현금흐름을 분석한다. 분석된 데이터는 비율로 수치화되어 경쟁사와 비교하고 장점과 약점을 도출한다.

2) 기회, 위기 분석

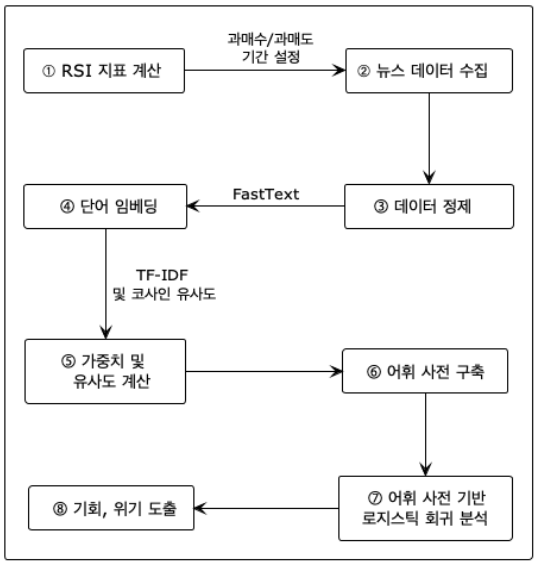

그림 1. 기회, 위기 분석 흐름도

그림 1은 FastText 모델을 이용한 기회, 위기 분석 흐름도를 보여준다.

Step 1: RSI(relative strength index) 지표를 이용하여 비교 기업의 과매수 및 과매도 시점들을 도출한다. 도출된 시점은 기회(과매수 시점)과 위협(과매도 시점) 분석의 기준 기간으로 설정된다.

Step 2: 설정된 기준 기간의 비즈니스 모델 및 경쟁사들과 관련된 뉴스 기사들을 크롤링함으로써 데이터들을 수집한다.

Step 3: 데이터 정제 과정을 통해 수집된 데이터들로부터 불필요하거나 노이즈를 포함하는 정보들을 제거한다. KoNLPy의 Mecab 파이썬 패키지를 이용하여 형태소를 분석하고 분절한다. 정제된 데이터들은 말뭉치(corpus)로 구성된다.

Step 4: FastText 알고리즘을 이용하여 말뭉치 내 단어들의 서브워드(subword)들을 임베딩(embedding)한다.

Step 5: TF-IDF 알고리즘과 코사인 유사도 기법을 이용하여 임베딩된 단어 벡터(vector)들의 가중치와 유사도를 계산한다.

Step 6: 계산된 가중치와 유사도 수치를 통해 기회, 위협 요인들의 SWOT 어휘 사전을 구축한다.

Step 7: 어휘 사전의 키워드들을 로지스틱 회귀 분석(logistic regression)의 입력 변수로 이용하여 뉴스 기사로부터 기회와 위협 요인을 추출한다.

Step 8: 학습된 로지스틱 회귀 모델을 기반으로 새로운 뉴스 데이터로부터 기회와 위기 요인을 도출한다.

제안된 방법론을 통해 창업 기업은 내부적 강점과 약점을 명확히 이해하고, 외부 환경에서의 기회와 위협을 파악할 수 있다.

Ⅳ. 결 론

본 연구에서는 창업 기업의 지속 가능성을 강화하기 위해 비즈니스 모델의 SWOT 분석을 수행하는 새로운 방법론을 제안하였다.

첫 번째로, 창업 기업 및 경쟁사의 공시자료를 통해 기업 가치를 수치화하고 내부 요인인 강점과 약점을 도출하였다.

두 번째로, RSI 지표를 기반으로 과매수/과매도 시점들을 파악한 후 해당 기간의 뉴스 기사들을 수집하였다. 그리고, FastText 모델을 기반으로 어휘 사전을 생성한 후 로지스틱 회귀 모델을 기반으로 외부 요인인 기회와 위협 분석을 수행하였다.

결론적으로 제안하는 방법론을 통해 창업 기업의 비즈니스 전략 수립에 기여하고, 의사결정 과정에서의 효율성과 정확성이 향상될 수 있을 것으로 기대한다.

References

[1] S. J. Park, “A case study on SWOT analysis for efficient use of BMC (focused on the case of Baekseolchimhyang),” *Master’s Thesis, Pusan National University*, Aug. 2018.

[2] Financial Supervisory Service. Data analysis, retrieval and transfer System (DART) [Internet]. Available: https://dart.fss.or.kr.

[3] H. Kang, J. Yang, “Performance comparison of Word2vec and fastText embedding models,” *Journal of Digital Contents Society*, Vol. 21, No. 7, pp. 1335-1343, Jul. 2020.

[4] D.-S. Park, H.-J. Kim, “A proposal of join vector for semantic factor reflection in TF-IDF based keyword extraction,” *The Journal of Korean Institute of Information Technology*, Vol. 16, No. 2, pp. 1-16, Feb. 2018.