Drivers and Barriers to Commercialization: An NLP Approach on Grant-Supported Startups

Beyda Tasar *
Department of Mechatronics Engineering, Firat University, Elazig, Turkiye

Talha Alperen Halisdemir
Department of Economics and Administrative Sciences, Fırat University, Elazıg, Turkiye

Izzet Tasar
Department of Economics and Administrative Sciences, Fırat University, Elazığ, Turkiye

Abstract

This study aims to develop and evaluate an automated natural language processing (NLP) framework specifically designed to extract commercialization enablers and barriers from interviews conducted with companies established through grant support programs. The research addresses the absence of domain-specific, language-adaptive NLP pipelines tailored to entrepreneurship and commercialization data in the Turkish context. By combining context-aware sentiment analysis with thematic classification, the proposed system is capable of processing large-scale, unstructured spoken content in Turkish in an efficient and interpretable manner.

The processing pipeline begins with the transcription of multi-format audio and video recordings using the OpenAI Whisper model, selected for its high accuracy in low-resource languages. The collected dataset comprises 48 semi-structured interviews with founders of grant-supported companies, totaling approximately 30 hours of speech, with each interview lasting between 30 and 45 minutes. These interviews span a diverse range of sectors, including technology, manufacturing, and services, thereby capturing a broad spectrum of commercialization experiences. The transcription stage achieves an average word error rate of 6.8%, ensuring a high-quality textual basis for subsequent NLP tasks.

Following transcription, the framework applies sentence segmentation and sentiment classification into positive, negative, and neutral categories. Sentiment detection is powered by a rule-enhanced, lexicon-based approach augmented with domain-specific positive and negative keyword lists that reflect commercialization-related contexts. This enables fine-grained identification of supportive factors—such as funding access, technical assistance, and collaborative partnerships—as well as obstructive factors like bureaucratic processes, market entry challenges, and regulatory uncertainty. Contextual extraction further enriches the analysis by capturing the preceding and following sentences surrounding each sentiment-bearing statement, preserving narrative coherence and improving interpretability.

The system also integrates keyword frequency analysis, stopword-filtered extraction of high-frequency terms, and batch processing capabilities to handle multiple interview files in a single run. Structured outputs in CSV and Excel formats facilitate direct integration into policy evaluation and business intelligence workflows. Benchmark tests demonstrate that the entire dataset can be transcribed and classified in under four hours on a standard GPU, indicating strong scalability.

The Turkish language introduces unique NLP challenges, including agglutinative morphology, complex suffixation, and flexible word order. These challenges are addressed through morphological normalization and stemming in the pre-processing stage, ensuring accurate lexicon matching and thematic tagging. The results confirm that the framework can effectively capture both enabling and hindering factors in the commercialization process, providing actionable insights for policymakers, investors, and academic researchers seeking to strengthen grant-funded entrepreneurial ecosystems. While the current version is primarily lexicon-based, the modular architecture allows for future integration of transformer-based contextual embeddings, expansion of thematic categories, and adaptation to other languages and domains.

Keywords

Commercialization, sentiment analysis, thematic tagging, NLP, Whisper,Turkish language processing, grant-supported companies.