Research Paper

Zuxing Wu, a1816653

The primary research question is: **Identifying and Comparing Emerging Technical Trends Across Platforms Using Keyword Extraction**

4.2 Data Collection

To comprehensively identify and compare emerging technical trends in robotics, we collected data from four major sources: GitHub, Reddit, Stack Overflow, and arXiv.

GitHub: We used the PyGithub library to search for repositories created after 2019-12-31 with keywords such as "robot", "robotic", "generative ai", "emotional intelligence", and "human-robot" in their name, description, or README. For each repository, we extracted and cleaned the README content, then applied the YAKE keyword extraction algorithm to identify relevant technical terms.

Reddit: Using the PRAW API, we searched Reddit posts from all subreddits with the same set of keywords. For each post, we combined the title and selftext, cleaned the text, and extracted keywords using YAKE. Duplicate posts were filtered out using their unique IDs.

Stack Overflow: We queried the Stack Exchange API for questions containing "robot" or "robotic" in the title, body, or tags, posted after 2019-12-31. The question titles and bodies were combined and processed with YAKE to extract technical keywords.

arXiv: We used the arXiv API to retrieve abstracts of recent papers matching the same set of keywords. Abstracts were deduplicated and processed with YAKE to extract top keywords.

For all platforms, we applied consistent data cleaning steps, including removing HTML tags, markdown links, and common stop words. The process was automated and parameterized (e.g., number of keywords, n-gram size) to ensure reproducibility. This multi-source approach enabled a broad and representative collection of technical terms and trends in the robotics domain.

4.3 Data Analysis

To analyze the collected data, we applied the YAKE keyword extraction algorithm to each dataset. YAKE is a lightweight, unsupervised method that identifies keywords based on their statistical properties in the text. We configured YAKE with parameters suitable for our datasets, such as n-gram size and maximum number of keywords per document.

The extracted keywords were then aggregated across all sources to identify common technical terms and trends. We performed a comparative analysis of the keyword distributions from each platform, focusing on the frequency and co-occurrence of terms. This analysis allowed us to identify emerging trends in robotics and compare their prevalence across different platforms.