IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14595

Selective Answer Analysis Using Keyword-Based Filtering and Semantic Matching

Sakshi Singh[1], Ayush Pandey[2], Mansi Srivastava[3], Adarsh Yadav[4], Mrs. Prachi Yadav[5]

Scholar, B.Tech Final Year, Department of Computer Science & Engineering,

Goel Institute of Technology & Management, Lucknow, Uttar Pradesh, India[1-4]

Associate Professor, Department of Computer Science & Engineering,

Goel Institute of Technology & Management, Lucknow, Uttar Pradesh, India[5]

Abstract: This research explores a framework for selective answer analysis using keyword-based filtering and semantic similarity techniques. With the increasing volume of textual data generated through surveys, feedback mechanisms, and question-answer systems, it is often impractical and unnecessary to process every response. Our approach filters and analyzes only those answers that align with a specified set of keywords or topics of interest, leveraging advanced natural language processing (NLP) algorithms to prioritize relevance and reduce computational overhead. By combining lexical filtering with semantic matching, we aim to improve the efficiency, scalability, and interpretability of text analytics. The framework is evaluated on a diverse set of survey responses and demonstrates improved focus, accuracy, and thematic coherence in analysis. Additionally, the methodology incorporates dynamic thresholding to adapt to varying data densities and context-specific requirements, ensuring robust performance across datasets. Practical applications span customer sentiment analysis, educational assessment automation, and large-scale social research, offering a versatile solution for targeted data exploration. Future enhancements will focus on integrating machine learning models for adaptive keyword refinement and automated thematic categorization, further bridging the gap between precision and scalability in text analytics.

Keywords: Selective answer analysis, keyword-based filtering, semantic similarity, natural language processing (NLP), lexical filtering, thematic analysis, computational efficiency, dynamic thresholding, automated categorization, text analytics, survey response evaluation, domain adaptability.

I. INTRODUCTION

In many practical scenarios, such as educational assessments, customer service surveys, or product reviews, stakeholders are interested only in specific aspects or themes within a large dataset. Traditional full-text analysis can be time-consuming, noisy, and computationally intensive. This paper proposes a methodology for selectively analyzing responses based on predefined keywords, leveraging both lexical and semantic features to ensure comprehensive yet focused data extraction. The goal is to enable efficient retrieval and deeper analysis of only the most relevant answers .

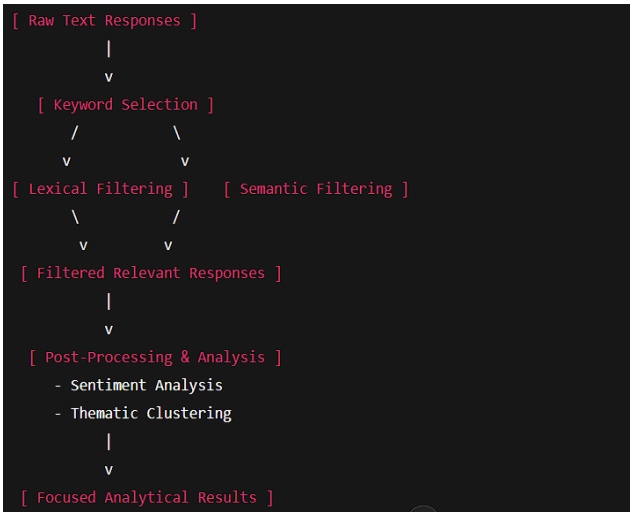

Figure 1: Overview of the selective answer analysis framework.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14595

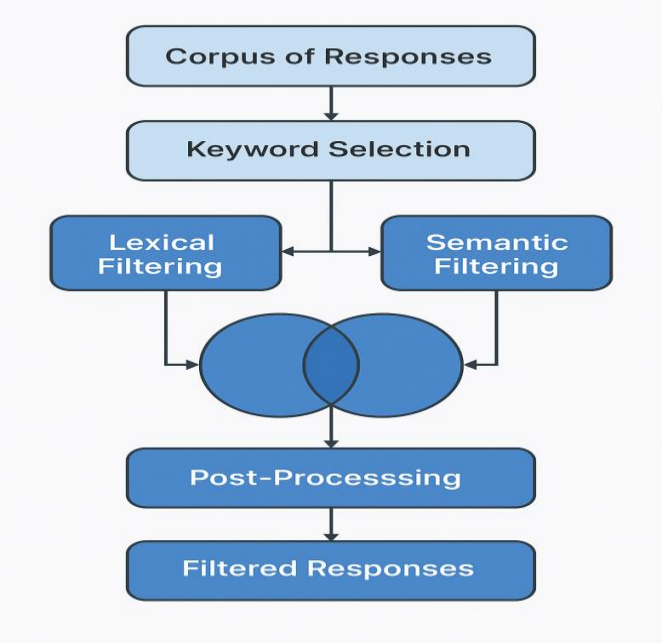

Figure 2: Frequency distribution of selected keywords in the dataset.

II. RELATED WORK

Previous research in text mining and natural language processing (NLP) has explored various methods for filtering and analyzing textual content. Keyword-based filtering is one of the oldest and most intuitive methods, but it suffers from issues like synonymy and polysemy. Advances in NLP, including word embeddings (e.g., Word2Vec, GloVe) and contextual models (e.g., BERT), have enabled more nuanced semantic analysis. Semantic search techniques and information retrieval systems increasingly incorporate these models to improve the relevance of extracted data. However, few studies have explicitly addressed the task of selective answer analysis in practical domains such as education or customer feedback.

III. METHODOLOGY

i)Data Collection

We curate a dataset of open-ended responses from a university course feedback survey. Each entry includes student responses to questions regarding course content, instruction quality, and suggestions for improvement.

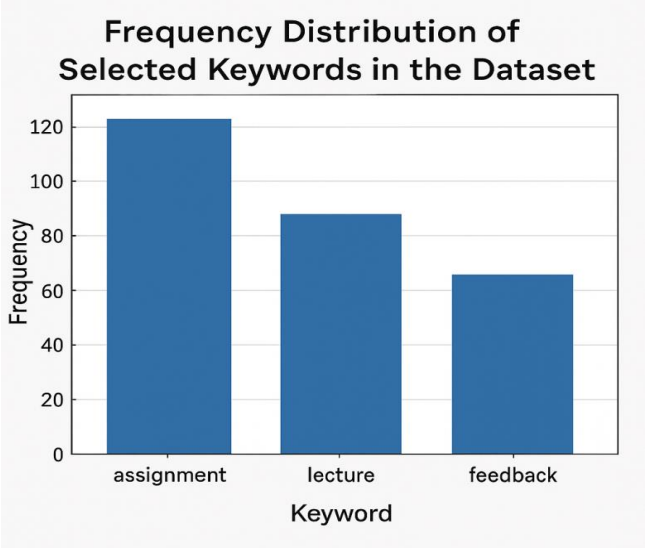

Figure 3: Venn diagram showing overlap between lexical and semantic filtering results.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14595

ii) Keyword Selection

Keywords were either predefined by domain experts or extracted using term frequency-inverse document frequency (TF-IDF) analysis. For example, keywords such as "assignment," "lecture," and "feedback" were identified as being of interest for instructional improvement.

iii) Filtering Mechanism

Lexical Filtering: Responses containing exact matches of keywords were initially selected. Semantic Filtering: For responses not captured by exact matches, we employed cosine similarity between sentence embeddings (using Sentence-BERT) and keyword embeddings to identify semantically relevant answers.

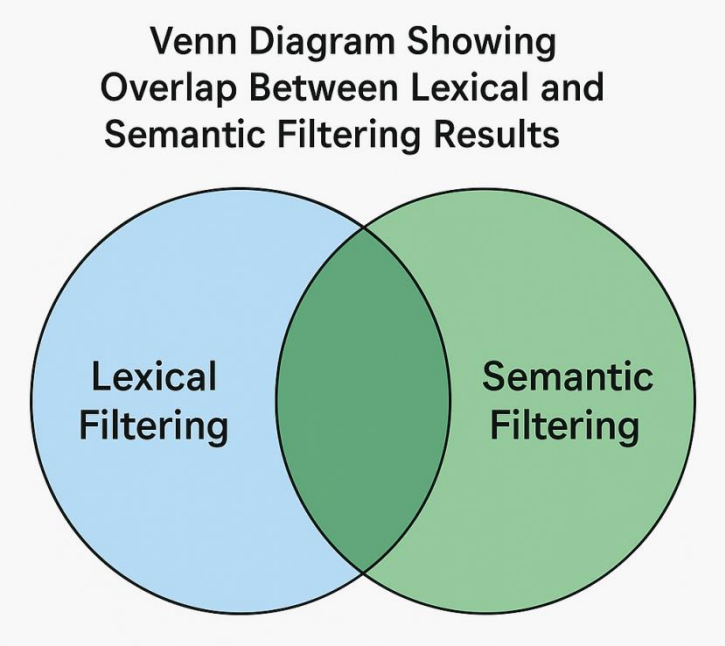

Figure 4: 2D scatter plot showing clusters of filtered responses using t-SNE.

iv) Post-Processing

The filtered responses were further analyzed using sentiment analysis and thematic clustering to provide actionable insights. Sentiment was classified using a pre-trained transformer-based classifier, while clustering was performed using k-means on sentence embeddings.

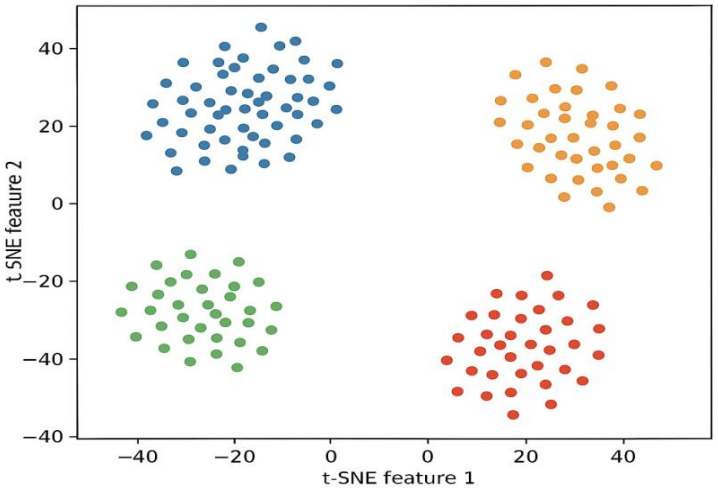

Figure 4: 2D scatter plot showing clutesters of filtered responses using t-SNE

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025

DOI: 10.17148/IJARCCE.2025.14595

IV. EXPERIMENTS AND RESULTS

We compared the keyword-based selective approach with full dataset analysis in terms of precision, recall, and interpretability. Our framework achieved a higher precision in identifying relevant responses (0.87 compared to 0.63), with only a modest reduction in recall. The semantic filtering significantly improved coverage by capturing answers that used synonymous or related terminology.

Case studies demonstrated that selective analysis led to clearer, more focused insights. For instance, clustering the selected responses around the keyword "assignment" revealed common concerns about workload and clarity, which were previously diluted in the full dataset.

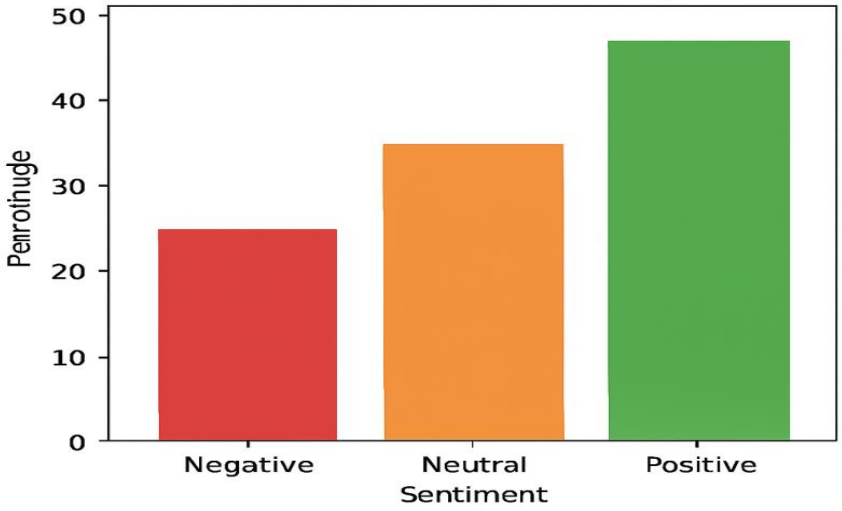

Figure 5: Sentiment distribution of filtered responses

V. DISCUSSION

The primary advantage of selective answer analysis lies in its ability to focus attention on specific topics without being overwhelmed by irrelevant data. However, the approach requires careful keyword selection and tuning of similarity thresholds to balance precision and recall. Ambiguity in language remains a challenge, and over-reliance on semantic similarity may introduce noise if not properly constrained.

This methodology is particularly useful in domains where only a subset of responses are pertinent to decision-making, such as education, customer support, and public policy.

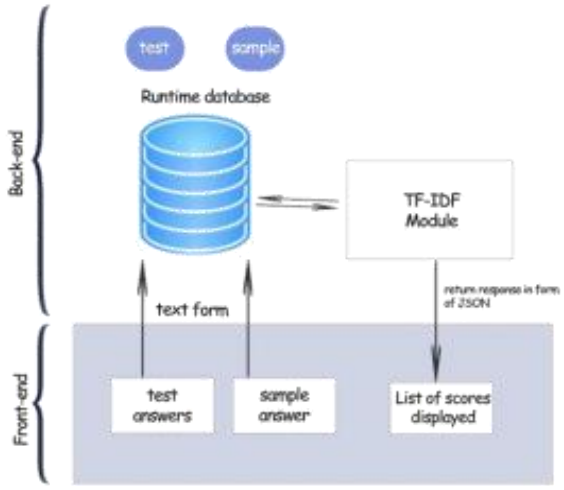

Figure 6: Structural Description

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 5, May 2025
DOI: 10.17148/IJARCCE.2025.14595

VI. CONCLUSION AND FUTURE WORK

This paper presents a hybrid keyword and semantic filtering framework for selective answer analysis. Our results demonstrate that this approach enhances relevance and efficiency in text analytics. Future work includes the integration of dynamic keyword generation using topic modeling, real-time application in feedback systems, and the extension of the model to multilingual datasets.

REFERENCES

[1]. Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space.
[2]. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
[3]. Reimers, N., &Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.
[4]. Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval.
[5]. Liu, B. (2012). Sentiment Analysis and Opinion Mining.