Mining App Reviews for User Feedback Analysis in Requirements Engineering: A Project Report

Quim Motger[1,*], Marc Oriol[1], Max Tiessler[1], Xavier Franch[1] and Jordi Marco[2]

[1]Department of Service and Information System Engineering, Universitat Politècnica de Catalunya
[2]Department of Computer Science, Universitat Politècnica de Catalunya

Abstract

Mining app reviews has emerged as a valuable practice in requirements engineering, providing insights into feature usage trends, user satisfaction, and emerging software issues. While recent advances in natural language processing have enhanced review analysis, challenges persist in feature extraction, sentiment ambiguity, and the scalability of automated methods, among others. This project report presents our research efforts in app review mining, focusing on methodological, software-based, and data-driven contributions. We explore both supervised and unsupervised learning approaches, leveraging large language models for key tasks such as feature identification, competition analysis, and emotion extraction. Additionally, we develop open-source tools and datasets to support reproducibility and adoption of our methods. Our findings highlight the potential of large language models in automating user feedback analysis while identifying gaps that require further research, particularly in addressing model reliability and evaluation challenges.

Keywords

app review mining, requirements engineering, user feedback analysis, natural language processing, large language models, feature extraction, competition analysis, emotion extraction

1. Introduction

Over the past decade, opinion mining has become an integral part of the software development lifecycle [1]. It is widely applied in multiple stages such as requirements elicitation [2], design specification [3], team management [4] and quality assurance [5]. In requirements engineering, crowdsourced repositories - such as issue tracking systems [6], social networks [7], and app stores [8] - offer a wealth of document-based knowledge, enabling the application of state-of-the-art, data-intensive natural language processing (NLP) methods. In the context of app store mining, these repositories present valuable research opportunities and practical benefits, including identification of emerging issues [8], requirements elicitation and prioritization [9], and opinion-driven software maintenance and evolution [10].

Despite significant advancements, several challenges persist in app review mining, especially with the rise of large language models (LLMs). From a data perspective, the lack of open-source datasets and standardized knowledge bases limits reproducibility and benchmarking [11], particularly in mitigating hallucinations and error-prone responses from LLMs. From a methodological perspective, attention-based mechanisms and transformer architectures have improved the analysis of crowdsourced, user-generated content [12]. However, several challenges such as managing context [13], handling sentiment ambiguity [1], and ensuring computational efficiency [14] still remain. From an evaluation perspective, improving precision and recall in tasks such as feature extraction and polarity analysis [15] remains crucial for the successful adoption of these techniques in industrial settings.

In this context, this paper presents a project report on our research group's contributions to mobile app review mining. Conducted as a non-funded initiative, our work focuses on three key objectives: (1)

In: A. Hess, A. Susi, E. C. Groen, M. Ruiz, M. Abbas, F. B. Aydemir, M. Daneva, R. Guizzardi, J. Gulden, A. Herrmann, J. Horkoff, S. Kopczyńska, P. Mennig, M. Oriol Hilari, E. Paja, A. Perini, A. Rachmann, K. Schneider, L. Semini, P. Spoletini, A. Vogelsang. Joint Proceedings of REFSQ-2025 Workshops, Doctoral Symposium, Posters & Tools Track, and Education and Training Track. Co-located with REFSQ 2025. Barcelona, Spain, April 7, 2025.

*Corresponding author.
joaquim.motger@upc.edu (Q. Motger); marc.oriol@upc.edu (M. Oriol); max.tiessler@upc.edu (M. Tiessler); xavier.franch@upc.edu (X. Franch); jordi.marco@upc.edu (J. Marco)
0000-0002-4896-7515 (Q. Motger)

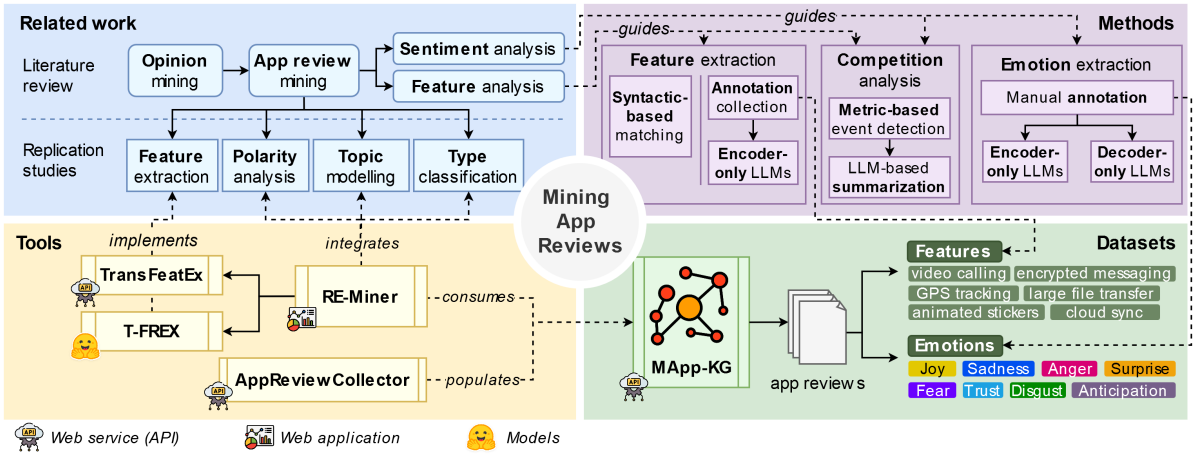

Figure 1: Project summary overview

synthesizing the state of the art in app review mining, (2) identifying research gaps, particularly in feature and sentiment analysis tasks, and (3) designing, developing, and distributing methodological, software-based, and data-driven contributions. We adhere to open science principles to ensure replicability and reusability. Ultimately, our goal is to advance the field by providing systematic insights and supporting practitioners and researchers in leveraging app reviews for software engineering, with a particular focus on requirements engineering.

2. Project Background

The Software and Service Engineering Research Group[1] (GESSI) at Universitat Politècnica de Catalunya (UPC-BarcelonaTech) specializes in software and requirements engineering, with a focus on software quality, architecture, AI-driven software engineering, and service-oriented computing. Within this scope, the group explores NLP for Software Engineering (NLP4SE), advancing tasks like requirements specification, knowledge base design, and feedback analysis. Recent efforts, driven by a dedicated PhD thesis [16], have contributed to large-scale initiatives such as OpenReq[2], which enhanced requirements analysis and stakeholder decision-making. More recently, GESSI has joined HIVEMIND[3], developing an LLM-based multi-agent framework for requirements elicitation, design specification, and task management. This project explores techniques like fine-tuning, prompt engineering, retrieval-augmented generation (RAG), and human-in-the-loop strategies to create a semi-automated ecosystem for intelligent software development.

The project reported in this paper is an independent, non-funded initiative led by our research group. While it aligns with OpenReq's outcomes and HIVEMIND's goals, it operates separately, focusing on complementary research directions. The project began in 2023 and is expected to conclude by 2026.

3. Project Summary

Figure 1 depicts a summarized overview of the scope and main contributions of this project. Specifically, we focus on four types of contributions: analysis of related work, methodological contributions, software-based tools and datasets or data-based artefacts (see our Data Availability Statement).

[1] https://gessi.upc.edu/en/
[2] https://cordis.europa.eu/project/id/732463
[3] https://cordis.europa.eu/project/id/101189745

3.1. Related Work

The main goal of our analysis of related work is to elicit existing research gaps and opportunities stemmed from the emergence of novel NLP methods and models. Specifically, our work is focused on:

- **Literature review**. Stemming from existing literature reviews in opinion mining [1] and app review mining [15], we have conducted literature reviews for the tasks of feature analysis and sentiment analysis. These have focused on: (1) analysis of related work for automated analysis; (2) software-based contributions; and (3) elicitation of remaining research gaps. The results are published in different publications (see Sections 3.2 and 3.3).
- **Replication studies**. Our research focus on specific contributions for relevant app review descriptors such as features (e.g., *send message*, *share private list*, *GPS navigation*...) and emotions (e.g., *Joy*, *Sadness*, *Anger*, *Fear*...) expressed by user feedback. Furthermore, in the analysis of related work, we have identified several app review descriptors which are relevant for multiple software engineering tasks. To complement and assist as enriched data in our tools (see Section 3.3), we have conducted several replication studies based on original publications to build automatic extraction services for review mining. These include: (1) polarity analysis (i.e., *positive*, *negative*), (2) topic modelling (e.g., *usability*, *security*, *aesthetics*, *learnability*), and (3) type classification (i.e., *bug report*, *feature request*, *app praise*, *user experience*).

3.2. Methods

The main goal of our methodological contributions in app review mining is to leverage modern NLP methods and models, with a particular emphasis on LLMs, to enhance the accuracy of traditional tasks such as feature extraction and emotion detection in user reviews. We explore both unsupervised and supervised learning approaches, placing significant focus on the collection and generation of open-source, annotated datasets tailored for specific review mining tasks. Specifically, our key contributions include:

- **Feature extraction**. We originally designed a syntactic-based pattern matching approach using encoder-only LLMs to extract syntactic annotations and extract traditional patterns from related work referring to feature mentions [17]. As an extension, we designed a proposal to leverage crowd-sourced annotations of features from app reviews to fine-tune encoder-only LLMs for the token classification task [18].
- **Competition analysis**. We have explored the potential of exploiting traditional app review mining descriptors (see Section 3.1) with generative AI to detect potential threats and opportunities (i.e., events) raised by user feedback within a specific app market segment [19]. Our proposal leverages generative LLMs to summarize and efficiently report such events to practitioners, assisting them into app market analytics and decision-making tasks such as requirements prioritization and release planning.
- **Emotion extraction**. As ongoing work, we are working on the annotation of mobile app reviews with a taxonomy of human emotions. We have selected a taxonomy of emotions and we have developed a set of guidelines to adapt such taxonomy to the app review domain. Furthermore, we have collaboratively worked to annotate a large dataset of app reviews with such emotions. We have monitored this process to identify challenges within the emotion extraction domain. Finally, we plan to extend this work by assessing automated extraction methods based on our dataset, including supervised learning methods fine-tuning encoder-only LLMs and few-shot approaches leveraging decoder-only LLMs (see Section 4).

3.3. Tools

In alignment with Open Science principles, all of our methodological contributions are embedded into software-based artifacts and distributed as open-source code under the GPL-3.0 license on GitHub repositories. Below, we highlight the most relevant tool-based contributions from our project:

- **AppReviewCollector**. A web-based service that combines API consumption with web scraping techniques to search and collect metadata and documents (e.g., summaries, descriptions, changelogs, reviews) from mobile apps. The current version includes app stores (e.g., Google Play), search engines (e.g., AlternativeTo[4]), and sideloading repositories (e.g., F-Droid[5]).
- **TransFeatEx**. A web-based service implementing the syntactic-based feature extraction method. The tool allows full parameterization and customization of the syntactic pipeline, including pre-processing configurations, syntactic patterns, and polarity-based filtering of app reviews.
- **T-FREX**. A collection of fine-tuned encoder-only LLMs (e.g., BERT, RoBERTa, XLNet) designed for token classification, specifically to identify feature-related named entities in app reviews. The models are available for download and inference on Hugging Face.
- **RE-Miner**. A web application designed to support app review mining tasks by integrating services, models, and datasets for various analysis processes. It incorporates our own contributions in feature and emotion extraction, along with existing approaches for polarity analysis, topic modeling, and type classification. The tool provides an adaptive architecture for deploying analytic tools as standalone microservices, allowing flexible customization of descriptors used in app review mining. Additionally, it includes a dashboard for analytical insights.

3.4. Datasets

Finally, our methodological and tool-based contributions have facilitated the collection and generation of datasets relevant to app review mining tasks. In alignment with Open Science principles, we actively promote their dissemination and use. Our work focuses primarily on three key data-based artifacts:

- **MApp-KG**. A knowledge graph cataloging 832 mobile apps across 46 app categories, containing 1,666 proprietary documents (such as descriptions and changelogs) and over 13 million user reviews. The knowledge graph is publicly available and adheres to the RDF schema, ensuring its reusability for future app review mining research.
- **Features from crowdsourced repositories**. A dataset of 23,816 user reviews with 32,443 feature mentions [18], collected from 468 mobile apps spanning 10 popular categories (e.g., *Productivity* and *Communication*). Feature annotations are sourced from a crowdsourced repository (AlternativeTo), which aggregates user-provided feature descriptions for mobile apps.
- **Emotions from manual annotation**. We are developing a dataset of mobile app reviews annotated with emotion labels derived from a consolidated emotion taxonomy, adapted to app review opinion mining tasks. Future work will focus on expanding and evaluating this dataset (see Section 4).

4. Research Plan

Our current research focuses on the analysis and automatic extraction of emotions from app reviews, aiming to establish emotions as a valuable descriptor in software and requirements engineering tasks. Additionally, we explore automated extraction methods and mitigation techniques to address challenges associated with specific emotions and limitations inherent in user feedback. Figure 2 provides an overview of the ongoing research plan in the emotion extraction research line.

- **Emotion annotation of app reviews**. As discussed in Section 3.2, we are finalizing the annotation of a large dataset of mobile app reviews using an adapted taxonomy of human emotions suited to the mobile app domain.
- **Annotation agreement analysis and evaluation of generative AI for annotation**. Based on feedback from multiple annotators, we have analyzed annotation agreement, identified key

[4] https://alternativeto.net/
[5] https://f-droid.org/es/

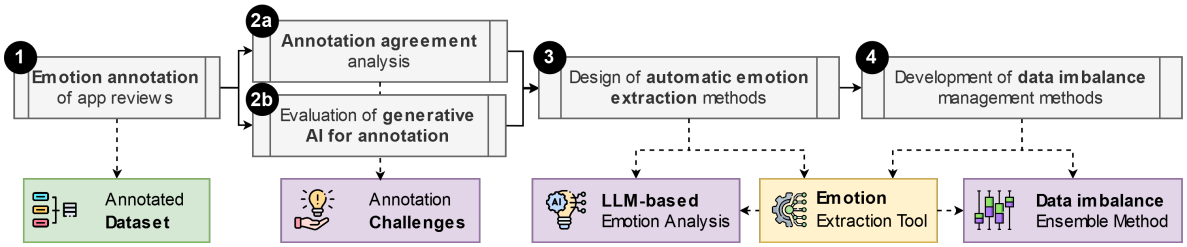

Figure 2: Research plan

challenges, and derived insights to improve the design of automated emotion extraction approaches. Additionally, we plan to assess the performance of generative AI models, such as GPT-4o, DeepSeek, and Mistral, as annotators by measuring inter-rater agreement with human annotations and evaluating their potential as a supplement or alternative to human annotators.

- **Design of automatic emotion extraction methods**. Using the annotated dataset and insights from previous tasks, we aim to compare the performance of encoder-only LLMs fine-tuned for text classification with decoder-only LLMs in a few-shot setting using guidelines and examples. The developed method will be integrated into the RE-Miner tool, expanding its analytical capabilities.
- **Development of data imbalance management methods**. Given the inherent imbalance in emotion distribution within our dataset, we plan to compare multiple strategies to mitigate data imbalance. Our goal is to improve the accuracy of underrepresented emotions while providing insights on effective techniques for handling imbalanced datasets in user feedback analysis.

5. Conclusions

Our research highlights the potential of app review mining in requirements engineering, using NLP and LLMs to automate tasks such as feature extraction, sentiment analysis, and competition monitoring. Alongside designing methodological contributions and developing open-source tools and datasets, we identified key challenges to address, including sentiment ambiguity and model reliability. Additionally, we recognize broader challenges, such as the practical integration of automated feedback analysis into software development. As future work, we plan to focus on refining emotion extraction methods and addressing data imbalances. By improving these methodologies and disseminating the generated datasets and tools, we aim to support more effective user feedback analysis, making it a practical asset for both researchers and practitioners.

Acknowledgments

With the support from the Secretariat for Universities and Research of the Ministry of Business and Knowledge of the Government of Catalonia and the European Social Fund. This paper has been funded by the Spanish Ministerio de Ciencia e Innovación under project / funding scheme PID2020-117191RB-I00 / AEI/10.13039/501100011033.

Data Availability Statement

All relevant contributions in this project, including source code repositories, data artifacts, and references to related publications, are indexed and documented in our replication package available at Zenodo.

References

[1] B. Lin, N. Cassee, A. Serebrenik, G. Bavota, N. Novielli, M. Lanza, Opinion mining for software development: A systematic literature review, ACM Trans. Softw. Eng. Methodol. 31 (2022).

[2] X. Liu, Y. Leng, W. Yang, C. Zhai, T. Xie, Mining android app descriptions for permission requirements recommendation, in: 2018 IEEE 26th International Requirements Engineering Conference (RE), 2018, pp. 147–158.

[3] B. Lin, F. Zampetti, G. Bavota, M. Di Penta, M. Lanza, Pattern-based mining of opinions in q&a websites, in: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), 2019, pp. 548–559.

[4] G. A. M. da Cruz, E. H. M. Huzita, V. D. Feltrim, Estimating trust in virtual teams - a framework based on sentiment analysis, in: Proceedings of the 18th International Conference on Enterprise Information Systems - Volume 1: ICEIS, 2016, pp. 464–471.

[5] H. Hu, S. Wang, C.-P. Bezemer, A. E. Hassan, Studying the consistency of star ratings and reviews of popular free hybrid android and ios apps, Empirical Software Engineering 24 (2019) 7–32.

[6] G. Destefanis, M. Ortu, S. Counsell, S. Swift, M. Marchesi, R. Tonelli, Software development: do good manners matter?, PeerJ Computer Science 2 (2016) e73.

[7] F. H. Khan, S. Bashir, U. Qamar, Tom: Twitter opinion mining framework using hybrid classification scheme, Decision Support Systems 57 (2014) 245–257.

[8] V. M. A. de Lima, R. M. Marcacini, Opinion mining for app reviews: Identifying and prioritizing emerging issues for software maintenance and evolution, in: Proceedings of the XXIII Brazilian Symposium on Software Quality, SBQS '24, 2024, p. 687–696.

[9] E. Guzman, W. Maalej, How do users like this feature? a fine grained sentiment analysis of app reviews, in: 2014 IEEE 22nd International Requirements Engineering Conference (RE), 2014, pp. 153–162.

[10] J. Dąbrowski, E. Letier, A. Perini, A. Susi, Mining user feedback for software engineering: Use cases and reference architecture, in: 2022 IEEE 30th International Requirements Engineering Conference (RE), 2022, pp. 114–126.

[11] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, H. Wang, Large language models for software engineering: A systematic literature review, ACM Trans. Softw. Eng. Methodol. 33 (2024).

[12] M. A. Hadi, F. H. Fard, Evaluating pre-trained models for user feedback analysis in software engineering: a study on classification of app-reviews, Empirical Software Engineering 28 (2023) 88.

[13] S. I. Ross, F. Martinez, S. Houde, M. Muller, J. D. Weisz, The programmer's assistant: Conversational interaction with a large language model for software development, in: International Conference on Intelligent User Interfaces, 2023, p. 491–514.

[14] N. Marques, R. R. Silva, J. Bernardino, Using chatgpt in software requirements engineering: A comprehensive review, Future Internet 16 (2024).

[15] J. Dąbrowski, E. Letier, A. Perini, A. Susi, Mining and searching app reviews for requirements engineering: Evaluation and replication studies, Information Systems 114 (2023) 102181.

[16] Q. Motger, Natural language processing methods for document-based requirements specification and validation tasks, Ph.D. thesis, Universitat Politècnica de Catalunya, 2024.

[17] A. Gallego, Q. Motger, X. Franch, J. Marco, Transfeatex: a NLP pipeline for feature extraction., in: Joint Proceedings of REFSQ-2023 Workshops, Doctoral Symposium, Posters & Tools Track, and Journal Early Feedback Track., 2023.

[18] Q. Motger, A. Miaschi, F. Dell'Orletta, X. Franch, J. Marco, T-FREX: A Transformer-based Feature Extraction Method from Mobile App Reviews, in: IEEE International Conference on Software Analysis, Evolution and Reengineering, 2024.

[19] Q. Motger, X. Franch, V. Gervasi, J. Marco, Unveiling competition dynamics in mobile app markets through user reviews, in: D. Mendez, A. Moreira (Eds.), Requirements Engineering: Foundation for Software Quality, Springer Nature Switzerland, Cham, 2024, pp. 251–266.