

Towards Enhancing Mobile App Reviews: A Structured Approach to User Review Entry, Analysis and Verification

Omar Haggag, John Grundy^a and Rashina Hoda

HumaniSE Lab, Department of Software Systems and Cybersecurity, Faculty of IT, Monash University, Australia

Keywords: Mobile Apps, User Reviews, Categorisation, ChatGPT, GPT-4, STGT, Tagging, Analysis, App Stores, Transparency.

Abstract: We propose an approach to address the shortcomings of current mobile app review systems on platforms such as the Apple App Store and Google Play. Currently, these platforms lack review categorisation and authentication of genuine user feedback, posing significant barriers for app developers and users. We propose an approach combining socio-technical grounded theory (STGT) and advanced natural language processing (NLP) tools such as GPT-4 to analyse user reviews, providing deeper insights into app functionalities, problems, and ultimately, user satisfaction. An interactive UI prototype is presented to demonstrate the use of structured, verified feedback. This includes a novel review submission process with categorisation/tagging and a "verified download" tag to ensure review authenticity. The goal of our approach is to enhance the app ecosystem by assisting developers in prioritising improvements and enabling users to make informed choices, encouraging a more robust and user-centric digital marketplace.


1 INTRODUCTION

In the dynamic and ever-expanding world of mobile applications, user reviews stand as an essential component in the digital ecosystem, linking app developers and users by providing transparent feedback on an app's performance, usability, and overall value (Vasa et al., 2012; Haggag et al., 2021). These insights are useful for developers, highlighting both strengths and areas needing improvement, thereby assisting their development strategies. For users, these reviews act as a guide through millions of available mobile apps, helping them in making informed decisions based on the shared experiences of others (Palomba et al., 2018). Not only do these reviews influence personal download, purchase and usage choices, but they also shape the evolution of apps to continually meet user expectations through their updates (Genc-Nayebi and Abran, 2017).

For app developers, user reviews are a significant feedback mechanism, revealing how their app performs in real-world scenarios, which might not be detected in testing environments (Li et al., 2018; Haggag, 2022). These reviews can reveal issues, from bugs to user experience problems. They also play a

significant role in analysing the success of updates and new features, influencing the app's developmental direction. On the user side, reviews are a significant resource for potential users, offering a more authentic look than promotional materials. Current users, through reviews, can share insights, contributing to the app's ongoing development and building community (Palomba et al., 2017).

However, a major challenge with the current review systems on platforms such as the Apple App Store and Google Play is the absence of review categorisation (Li et al., 2018). This complicates developers' ability to accurately and effectively analyse and respond to feedback, especially with reviews often covering multiple issues, spelling and grammar mistakes, and are sometimes submitted in different languages. Another significant concern is distinguishing genuine user reviews from fake ones submitted by people or generated by bots. Currently, there is no definitive way to indicate or verify if a review is genuine, which can skew the perception and reliability of the feedback (Martens and Maalej, 2019; Haggag et al., 2022a; Haggag et al., 2022b). Furthermore, for paid or subscription-based mobile apps without a system in place to confirm if a user has made a purchase or subscribed to services within the app, there's no guarantee that reviews reflect real customer experience.

^a  <https://orcid.org/0000-0003-4928-7076>

riences. This gap presents a high risk of fake reviews, where individuals who have not actually bought or engaged with the app can leave misleading feedback, potentially affecting the app's reputation and user decisions.

We propose the use of socio-technical grounded theory (STGT) to provide a structured and more comprehensive approach to analysing app reviews, offering insights into both the technological aspects of the app and the social context of user interactions (Hoda, 2021). By applying STGT, researchers and app developers can better understand patterns and themes in user reviews that go beyond simple functionality issues, understanding how user sentiments evolve with app updates or how social influences shape app perception (Hoda, 2021; Hoda, 2023; Fazzini et al., 2022). This analysis, empowered by text analysis tools using natural language processing such as GPT-4 combined with manual coding, can lead to a better understanding of the user-app relationship. Also, it enables developers to make user-centric enhancements and align the app more closely with user needs and preferences in this digitally interconnected world (Sanderson, 2023).

We conducted a study to better understand (i) the limitations of current reviewing mechanisms in app stores; (ii) the challenges faced by app developers, current and potential app users dealing with user reviews in the current structure; and (iii) key areas for improvement. The key contributions of this work include:

- analysis of how a structured, authentic review system can assist potential users in making more informed decisions, enriching their app selection and usage experience;
- designing an interactive UI prototype as a proof of concept highlighting the impact of organised, reliable user feedback on the app development cycle, particularly in terms of addressing user-specific issues and feature enhancement;
- creating a structured review submission methodology using categorisation/tagging, grounded in STGT principles, to streamline and optimise the extraction of useful feedback;
- introduction of "Verified Download" and "Verified Purchase" tags to enhance the credibility and authenticity of user reviews, ensuring that feedback is sourced from real users; and
- developing a tool prototype to illustrate how novel NLP tools such as GPT-4 combined with STGT can greatly improve the user review submission and analysis process.

2 MOTIVATION

Unstructured and Uncategorised Reviews: Figure 1 shows an example of the style of the current user interface of app reviews in the App Store and Google Play. The current mechanism for submitting user reviews on major platforms like the Apple App Store and Google Play has significant limitations for app developers, current users, and potential users alike (Jacob and Harrison, 2013; Martens and Maalej, 2019). The absence of categorisation in the review feedback system leads to a large, unstructured text of user opinions and experiences. For app developers, going through this large amount of data to extract actionable insights is a challenging task. Key issues and popular feature requests can be hidden among less relevant content, slowing down the response and resolution time and potentially leading to misguided priorities (Ciurumelea et al., 2017). Moreover, the lack of review structure often results in valuable feedback being overlooked or lost.

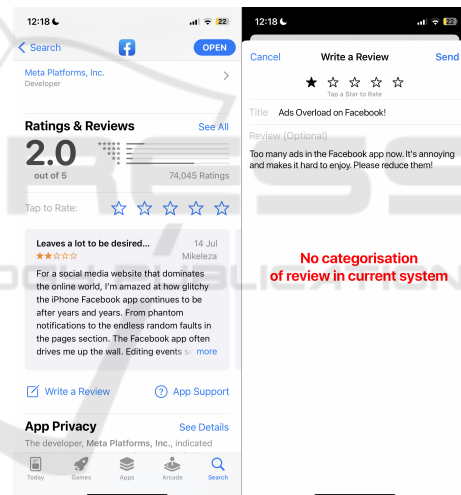


Figure 1: Current App Review UIs Lacking Categorisation.

App users looking to understand specific aspects of an app, such as its performance, usability, or particular features, must navigate through a large amount of general and irrelevant reviews (Vu et al., 2015). This process is time-consuming and can be overwhelming, worsening the overall experience and possibly leading to misinformed decision-making by the app designers and developers. The credibility of these reviews is another concern. With the prevalence of bot-generated reviews and the difficulty in knowing authentic user experiences from fake ones, users often struggle to know the true quality and reliability of an app (Caldeira et al., 2017; Martens and Maalej, 2019).

By categorising user review data into multiple

different themes or aspects, such as *usability*, *features*, *bugs*, and *user interface (UI)*, a more structured framework for both submitting, identifying, and analysing user feedback can be achieved. This would enable developers to quickly identify and prioritise areas that require attention, enhancing the efficiency and effectiveness of the development process. Potential app users could easily find the information most relevant to their interests or concerns, leading to a more satisfying and informed app selection process. Leveraging the latest NLP tools can also support more accurately capturing, categorizing and interpreting the various nature of user reviews.

Verifying Quality of Reviews: Existing app review systems lack mechanisms to verify whether user reviews originate from actual app downloads or confirmed purchases (Martens and Maalej, 2019). This raises concerns about the authenticity of the feedback, potentially enabling an increase in fake reviews or bot-generated content. Introducing a "Verified Download" status and possibly a "Verified Purchase" tag alongside user reviews could significantly mitigate these issues. Such verification processes would ensure that the feedback comes from users who have genuinely downloaded and interacted with the app, enhancing the credibility and value of the reviews for both developers and other users.

Review Submission Process: Existing research on user reviews in app stores predominantly focuses on qualitative and quantitative analyses of the content of the reviews, with minimal emphasis on enhancing the review submission process itself (Ciurumelea et al., 2017; Huebner et al., 2018; Li et al., 2017; Fu et al., 2013; Alqahtani and Orji, 2020). This misses the crucial aspect of how user reviews are collected and organised to ensure quality and timeliness. We want to provide a structured and systematic approach to the review capture and analysis process.

3 METHOD

We propose to leverage socio-technical grounded theory (STGT) and the natural language processing (NLP) capabilities of GPT-4 to aid in better categorisation of app reviews, and advocate for the use of Verified Download and Purchase tags. We aim for this method not just to enhance the clarity and relevance of user reviews, but also to increase the overall trustworthiness and usefulness of app reviews for all stakeholders.

3.1 User Reviews Classification Process Using STGT and NLP

Our proposed user review categorisation methodology for app stores is a two-phase system designed to enhance the utility and relevance of user feedback. The categorisation process outlined in Figure 2 represents a structured approach to managing user reviews on an app store. In the initial submission phase - step 1, users write and submit their reviews, which are then preprocessed using NLP techniques to extract key terms and sentiments in steps 1.1 and 1.2. Users can further tag their reviews with hashtags to highlight specific elements in step 1.3, after which the system, informed by STGT-based analysis, suggests relevant aspects for categorisation as in step 1.4. Users then have the opportunity to verify or adjust these suggestions before the review is added to the database, as in step 1.5.

Post-submission, reviews are tagged with "Verified Download" or "Verified Purchase" to indicate authenticity as in Step 2.1, and developers are notified of the new classified feedback in step 2.2. The published reviews in step 2.3 then become part of a continuous learning cycle, where the combined NLP system and STGT framework evolve based on emerging trends from new user feedback, ultimately allowing both users and developers to filter and leverage reviews more effectively in steps 2.4 and 2.5.

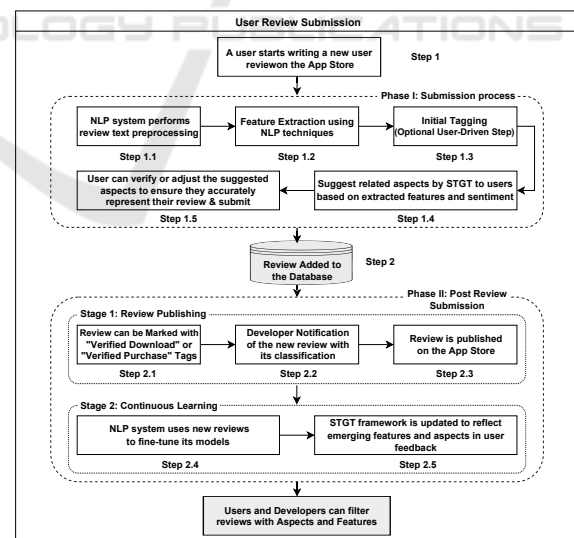


Figure 2: Proposed user reviews submission process.

The proposed classification process is initiated the moment a user starts writing a review. As they type, our NLP algorithm using GPT-4 will analyse the content in real-time, suggesting relevant themes and categories based on STGT-informed underlying models.

These suggestions are derived from a set of common 'seed' themes added into the model and previously identified themes within the user feedback pool, such as usability, functionality, performance, and customer service. This predictive categorisation will help users classify their reviews with the most relevant aspects, improving the structure and searchability of their feedback. However, our proposed system is designed to empower users with the freedom to select their own tags or create new ones, ensuring that the categorisation remains flexible and user-driven. When new user-defined aspects are introduced, they are fed back into the STGT framework, which is significant for capturing the evolving landscape of user experiences and expectations.

The STGT approach guides this adaptive process by providing a socio-technical lens, ensuring that both social aspects (like UX and satisfaction) and technical aspects (such as app functionality and bug reports) are captured and reflected in the evolving classification model. The theoretical framework acts as a backbone for understanding the complex interplay between the app's technical features and the social context of its use. Specifics of its application will inevitably require experimentation and revisions in practice.

Simultaneously, NLP algorithms work in the background to refine and expand the existing classification model. They process the natural language of the reviews, adapting to new slang, terminologies, and emerging issues. This dynamic approach guarantees that the classification system remains up-to-date with the latest trends and user concerns. Once a review is submitted, it undergoes further analysis to confirm the initial categorisation. The NLP system revisits the content, applying more rigorous text analysis techniques to ensure that the final categorisation aligns with the inductive analysis approach of the STGT framework applied to this context. This includes correlating the user-selected aspects with the model's suggestions and adjusting the categorisation model accordingly if needed.

Figure 3 illustrates the proposed interface in use, designed using Figma. Visual Cues and an accessible tagging system make it easy for users to contribute to the classification while expressing their feedback in a structured manner that directly informs app development and improves the UX for future app iterations. This will facilitate a clear, precise review classification process that is not only automated and efficient but also deeply rooted in the genuine needs and contributions of the app's user base. This dual approach ensures that the review system evolves in conjunction with user expectations and the app's development, leading to a robust mechanism for quality feedback and continuous improvement.

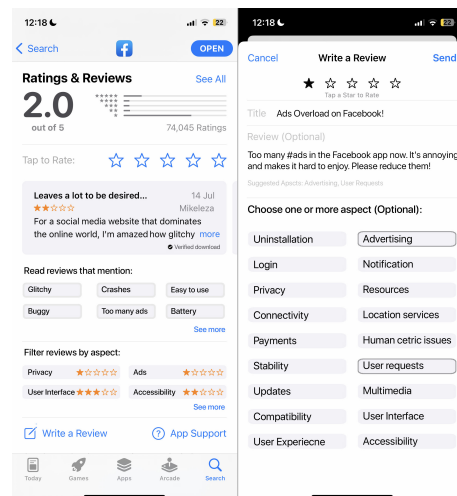


Figure 3: Proposed User Interface of Review Submission – see prototype here.

3.2 Implementation of "Verified Download" and "Verified Purchase" Tags

"Verified Download" and "Verified Purchase" tags alongside user reviews can improve the integrity and relevance of feedback on app stores. For "Verified Download", each time an app is downloaded, the app store's system records this event tied to the user's account ID. When the user decides to leave a review for the app, the system checks this record to confirm if a legitimate download has occurred. Upon successful verification, the user's review is automatically marked with a "Verified Download" tag, as shown in Figure 4. This badge of authenticity is then displayed next to the review on the app's storefront to inform potential users that the feedback comes from an actual user experience.

For "Verified Purchase", the system logs any in-app purchases or subscriptions linked to the user's account. When a review is posted, it cross-references this data to verify whether the reviewer has made a financial commitment to the app. If a purchase or subscription is verified, the review is marked with a "Verified Purchase" tag as shown in Figure 4. This process ensures that reviews reflecting the paid features of the app are easily distinguishable, providing prospective users with insights from those who have fully engaged with the app's offerings.

Technical implementation of these features will need to prioritise data privacy, ensuring that only essential data is used for verification purposes and protecting against the falsification of verification tags. The system should be capable of real-time updates, reflecting the verified status immediately after a

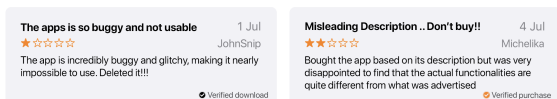


Figure 4: Reviews with verified download/purchase tags.

download or purchase occurs, regardless of the user’s device or firmware version. Moreover, the design of the “Verified” tags within the user interface will be clear to users but not disruptive to the overall UI of the app store’s review section. By adopting these enhancements, app stores will offer a review system that not only helps developers in obtaining genuine feedback but also enables users to recognise and trust the authenticity of reviews, facilitating more informed decisions regarding their app downloads and purchases.

3.3 Tool Prototype

We have developed a prototype tool to classify user reviews and suggest relevant aspects in real-time as the user types their feedback, as shown in Figure 5. At the core of this system is a suite of NLP techniques, primarily using a combination of Named Entity Recognition (NER) for extracting specific entities and aspects from the text and Sentiment Analysis to capture the emotional tone of the review. Leveraging the power of Transformer-based models, particularly GPT-4, the tool dynamically processes the input text to identify key themes and user sentiments. Unlike Bidirectional Encoder Representations from Transformers (BERT), which analyses text input bidirectionally but independently of the context for each word, GPT-4’s transformer architecture facilitates an understanding of each word in relation to the entire sentence structure, which significantly enhances the contextual relevance of aspect identification and sentiment interpretation. Concurrently, STGT analysis framework is employed to interlink the extracted entities and sentiments with broader socio-technical pre-defined aspects, providing users with intuitive aspect suggestions that reflect the context and content of their reviews. This feature not only enhances the review’s richness in detail but also aids in categorising the feedback for more actionable insights. The algorithmic workflow is fine-tuned through continuous learning, using up-to-date user review data to refine its predictive capabilities and ensure high accuracy and relevance in its suggestions.

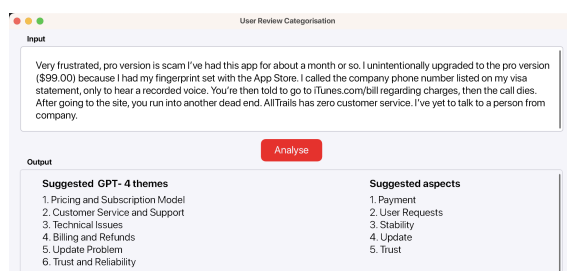


Figure 5: A screenshot of our prototype tool for user review classification using GPT-4 and STGT.

4 EVALUATION OF BENEFITS

Enhanced Feedback Relevance and Prioritisation: Developers benefit from receiving feedback that is both categorised and sentiment-analysed, which improves the focus on user concerns that are most critical. This enables developers to prioritise updates and features that will have the most significant impact on user satisfaction and app performance.

Quality Assurance and User Engagement Insights: The incorporation of “Verified Purchase” and “Download” tags assures developers that the feedback is sourced from users who have genuinely interacted with the app, providing a solid foundation for quality assurance. Additionally, understanding user engagement through the analysis of verified reviews informs developers about which features or updates resonate best with the users.

Resource Optimisation and Market Insight: The clear categorisation of feedback streamlines the review analysis process, allowing developers to allocate their resources more effectively to address bugs and develop new features. Furthermore, insights from sentiment analysis and STGT offer a deeper understanding of market reception, which can inform strategic business decisions.

Empowered User Feedback and Community Building: For current app users, the visibility of categorised and valued feedback empowers them to provide more detailed input, knowing that their concerns are recognised and acted upon. This not only encourages a richer dialogue but also fosters community spirit as users witness their collective voice influencing app evolution.

Informed Decisions and Time Savings for Potential Users: Potential app users gain the ability to make more informed decisions based on the categorised and verified reviews, which reflect real user experiences. The categorisation of reviews by key aspects like usability and performance means that potential users can quickly find the most significant information, sav-

ing time and aiding in a more efficient app selection process.

Trust in App Quality and Risk Reduction: The "Verified Purchase" tags signal a level of investment and satisfaction from existing users, enhancing trust in the app's quality for potential users. This, coupled with the insights from reviews of verified users, allows potential users to assess the risks associated with downloading or purchasing the app, ensuring they are more comfortable and confident in their choices.

5 NEXT STEPS

In our future work, we aim to implement the integration of STGT with NLP techniques of our review classification system, particularly focusing on optimising the precision of GPT-4 in sentiment and entity recognition to better capture and analyse user feedback. A significant expansion would be the adaptation of the system for direct integration of the classification tool with app development and feedback platforms, allowing for a smooth feedback loop that could directly influence app updates and feature enhancements. Additionally, we plan to explore the application of predictive analytics to preemptively identify user trends and enable proactive improvements to the app experience. This future work, prioritising algorithmic sophistication, cross-platform and multilingual support, and predictive capabilities, is expected to significantly advance the responsiveness and user-centeredness of app development practices.

ACKNOWLEDGEMENTS

Haggag and Grundy are supported by ARC Laureate Fellowship FL190100035.

REFERENCES

- Alqahtani, F. and Orji, R. (2020). Insights from user reviews to improve mental health apps. *Health informatics journal*, 26(3):2042–2066.
- Caldeira, C., Chen, Y., Chan, L., Pham, V., Chen, Y., and Zheng, K. (2017). Mobile apps for mood tracking: an analysis of features and user reviews. In *AMIA Annual Symposium Proceedings*, volume 2017, page 495. American Medical Informatics Association.
- Ciurumelea, A., Schaufelbühl, A., Panichella, S., and Gall, H. C. (2017). Analyzing reviews and code of mobile apps for better release planning. In *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 91–102. IEEE.
- Fazzini, M., Khalajzadeh, H., Haggag, O., Li, Z., Obie, H., Arora, C., Hussain, W., and Grundy, J. (2022). Characterizing human aspects in reviews of covid-19 apps.
- Fu, B., Lin, J., Li, L., Faloutsos, C., Hong, J., and Sadeh, N. (2013). Why people hate your app: Making sense of user feedback in a mobile app store. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1276–1284.
- Genc-Nayebi, N. and Abran, A. (2017). A systematic literature review: Opinion mining studies from mobile app store user reviews. *Journal of Systems and Software*, 125:207–219.
- Haggag, O. (2022). Better identifying and addressing diverse issues in mhealth and emerging apps using user reviews. pages 329–335.
- Haggag, O., Grundy, J., Abdelrazek, M., and Haggag, S. (2022a). Better addressing diverse accessibility issues in emerging apps: A case study using covid-19 apps.
- Haggag, O., Grundy, J., Abdelrazek, M., and Haggag, S. (2022b). A large scale analysis of mhealth app user reviews. *Empirical Software Engineering*, 27(7):196.
- Haggag, O., Haggag, S., Grundy, J., and Abdelrazek, M. (2021). Covid-19 vs social media apps: Does privacy really matter? pages 48–57.
- Hoda, R. (2021). Socio-technical grounded theory for software engineering. *IEEE Transactions on Software Engineering*, 48(10):3808–3832.
- Hoda, R. (2023). Technical briefing on socio-technical grounded theory for qualitative data analysis. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion)*, pages 344–345. IEEE.
- Huebner, J., Frey, R. M., Ammendola, C., Fleisch, E., and Ilic, A. (2018). What people like in mobile finance apps: An analysis of user reviews. In *Proceedings of the 17th international conference on mobile and ubiquitous multimedia*, pages 293–304.
- Iacob, C. and Harrison, R. (2013). Retrieving and analyzing mobile apps feature requests from online reviews. In *2013 10th working conference on mining software repositories (MSR)*, pages 41–44. IEEE.
- Li, X., Zhang, Z., and Stefanidis, K. (2018). Mobile app evolution analysis based on user reviews. In *New Trends in Intelligent Software Methodologies, Tools and Techniques*, pages 773–786. IOS Press.
- Li, Y., Jia, B., Guo, Y., and Chen, X. (2017). Mining user reviews for mobile app comparisons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–15.
- Martens, D. and Maalej, W. (2019). Towards understanding and detecting fake reviews in app stores. *Empirical Software Engineering*, 24(6):3316–3355.
- Palomba, F., Linares-Vásquez, M., Bavota, G., Oliveto, R., Di Penta, M., Poshyanyk, D., and De Lucia, A. (2018). Crowdsourcing user reviews to support the evolution of mobile apps. *Journal of Systems and Software*, 137:143–162.

- Palomba, F., Salza, P., Ciurumelea, A., Panichella, S., Gall, H., Ferrucci, F., and De Lucia, A. (2017). Recommending and localizing change requests for mobile apps based on user reviews. In *2017 IEEE/ACM 39th International Conference on Software Engineering (ICSE)*, pages 106–117. IEEE.
- Sanderson, K. (2023). Gpt-4 is here: what scientists think. *Nature*, 615(7954):773.
- Vasa, R., Hoon, L., Mouzakis, K., and Noguchi, A. (2012). A preliminary analysis of mobile app user reviews. In *Proceedings of the 24th Australian computer-human interaction conference*, pages 241–244.
- Vu, P. M., Pham, H. V., Nguyen, T. T., and Nguyen, T. T. (2015). Tool support for analyzing mobile app reviews. In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 789–794. IEEE.

