ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 10, Issue 12, December 2023

Systematic Literature Review on User Experience Analysis of Mobile Applications using Natural Language Processing Approach

Muhammad Zhafari Syah, I Gede Mujiyatna

Student, Department of Computer Science, Universitas Gadjah Mada, Sleman, Special Region of Yogyakarta, Indonesia
Lecturer, Department of Computer Science, Universitas Gadjah Mada, Sleman, Special Region of Yogyakarta, Indonesia

ABSTRACT: With the proliferation of mobile applications across diverse domains, ensuring an optimal user experience (UX) has become a critical concern for developers, researchers, and industry stakeholders. This systematic literature review (SLR) synthesizes existing research on the application of Natural Language Processing (NLP) techniques for the analysis and enhancement of user experience in mobile applications. The study aims to provide a comprehensive overview of the methodologies, challenges, and trends associated with leveraging NLP in the evaluation and improvement of mobile app UX. The primary focus is on studies that investigate user feedback, reviews, and other textual data related to mobile applications through NLP methodologies. In this paper, we have reviewed the existing techniques which encompass a broad range of NLP techniques, including sentiment analysis, opinion mining, and text summarization, among others.

KEY WORDS: Systematic literature review, User Experience, Mobile Application, Natural Language Processing, Machine learning, Deep learning

I.INTRODUCTION

In the current information era, user experience has emerged as a pivotal element that can significantly impact the success of digital products and services. The ability to understand, measure, and enhance UX is a critical concern for both researchers and practitioners in the field of human-computer interaction, as well as various other domains where technology interfaces with users [5]. As the volume of digital content and user interactions continues to grow, several studies have investigated the feasibility of the use of some methods in leveraging user-generated content to provide insights on the user experience of a given digital product or service [9, 11, 10, 12].
This systematic literature review seeks to provide a comprehensive overview of the state-of-the-art approaches, methodologies, and trends in the field of user experience analysis through the lens of Natural Language Processing.

II. SIGNIFICANCE OF THE SYSTEM

This literature review mainly focuses on how natural language approaches can be used to analyze the user experience of mobile applications. Methodology is presented in section III, section IV covers the experimental results of the study, and section V discusses the future study and Conclusion.

III. METHODOLOGY

First we are going to define our Population, Intervention, Comparison, Outcome, and Context (PICOC) using keywords as shown in Table 1. Table 1 provides a framework for a study aimed at understanding and improving user experience with mobile applications, with a specific focus on government mobile applications.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 10, Issue 12, December 2023

Table 1. PICOC Keywords

Points	**Keywords**
Population	User experience analysis, user review analysis
Intervention	Textual analysis, traditional analysis, user research
Comparison	Natural language processing, machine learning, deep learning
Outcome	User experience pain points, user experience issues, user experience insights
Context	Mobile applications, government mobile applications

From the Table 1 definition of our keywords, we can now define our search strings which are:

- Search string 1: (“user experience analysis” OR “user review analysis”) AND (“natural language processing” OR “machine learning” OR “deep learning”) AND “mobile application”
- Search string 2: (“user experience analysis” OR “user review analysis”) AND (“natural language processing” OR “machine learning” OR “deep learning”) AND “mobile application” AND “Indonesia”

The 2 search strings are similar with the main difference being that ‘Search string 2’ has the keyword “Indonesia” at the end of it to focus on finding research that is either written by Indonesian authors, or is analyzing Indonesian data, objects, or entities.

Next we define our research questions. We do this by looking at our original goal which is to provide an overview of the state-of-the-art approaches, methodologies, and trends in the field. From there we generated 5 research questions that are relevant to our goal which are:

1. What methods are currently used to analyze the user experience of mobile applications?
2. What is the impact of natural language processing on textual user experience analysis of mobile applications?
3. What are the challenges of using natural language processing and textual analysis in user experience analysis of mobile applications?
4. What are the natural language processing methods that are appropriate for textual user experience analysis of mobile applications?
5. What kind of results are synthesized from textual user experience analysis of mobile applications?

The next step is to select our digital libraries. Here we select Scopus, ScienceDirect, Google Scholar, and IEEE as our digital libraries where we will be conducting our collection of papers that is going to be reviewed. The digital libraries are selected because they are the digital libraries that are available to be accessed that are also relevant to the field that we are currently researching. Next we define our inclusion and exclusion criteria. We are only going to review articles that are published from 2013 and up, are in the English or Indonesian language, are not from books, and that have content that at least answers 2 of our research questions. We also assess the quality of the papers by seeing whether or not they discuss problems in the case of validity of their results and whether or not they have a clear definition or description of their goals, purposes, motivations, objectives, and/or research questions.

From the previously mentioned criterias, we found the following result when we enter our first and second search string into the public libraries as shown in Table 2. Table 2 provides a comparative analysis of the number of results retrieved from various digital libraries using two distinct search strings. The data within the table is critical for assessing the comprehensiveness and responsiveness of these libraries to specific queries, which can be instrumental for researchers when selecting resources for conducting a literature review or gathering data for their studies.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 10, Issue 12, December 2023

Table 2. Number of results from digital libraries

Public library	Search string 1	Search string 2
Scopus	1 result	0 result
ScienceDirect	16 results	0 result
Google Scholar	131 results	13 results
IEEExplore	219 results	21 results

After shifting through the results, we found 12 articles that meet our predefined criteria. From those 12 articles, we are going to try to find answers for our research questions and extract other information such as technology, framework, or platform used, the application field, gaps and challenges, findings, evaluation methods.

IV. RESULTS

A. Data Collection Methods

To use NLP models to do user experience analysis on mobile applications, researchers would need a lot of text data relating to the user experience of the application that is being reviewed. This kind of text data can be obtained from conducting surveys or online questionnaires targeted at the users of the application which has been done on previous research [1]. Although, the amount of data that can be collected through those online surveys and questionnaires only ranges around 300 to 500 respondents. A large-scale research on a single category on the Google Play Store shows that 34.63% of applications in that category alone have over 1 million downloads [4]. Thus we can conclude that surveys or online questionnaires may not be able to capture the whole picture of the user experience of a given application due to the large gap between the number of respondents to the number of actual users.

To obtain large amounts of text data that can better capture the user experience of a given application, many researchers have utilized the use of web crawlers or web scrapers to extract user reviews that are available on online application stores such as the Apple App Store, the Google Play Store, and the Microsoft Store [2, 11, 12]. Despite that, there are still some researchers that manually analyze the user experience of mobile applications by having a team of selected raters go through an application rating process which is then complemented by the use of textual analysis with some NLP method [3].

B. Pre-processing and Text Cleaning

When the data collection process is done, the next step to doing textual user experience analysis is to clean up the data. This is especially important if the data was obtained by the use of web crawlers or web scrapers from application stores accessible by mobile devices. User reviews that are written by the users of mobile applications usually contain misspelled words, abbreviations, and acronyms [7]. This problem is usually handled by using a custom dictionary that would correct the misspelling, abbreviations, and acronyms into their actual word which is also tailored to the language of the review text [8, 11]. Other than abbreviations, misspelling and acronyms, NLP models often have difficulty in detecting words that have prefixes or suffixes such as '-ing' suffixes in english words like 'landing' and 'maintaining' and 'ber-' prefixes in words like 'berkerja' and 'berbunyi' in the case of text in that are in Indonesian. Researchers handle this by using either stemming [10] or lemmatization [4]. Stemming reduces the words that have suffixes and prefixes by following a set of rules. An example of stemming is reducing the word 'trading' into 'trad' by removing the '-ing' suffix from the word. On the other hand, lemmatization relies on a dictionary that converts words that have a similar meaning into a single word. An example of lemmatization is turning the word 'went' and 'gone' into a word that has the same meaning which is 'go'. Other common text cleaning includes the removal of special symbols and stopwords and the conversion of the text into a uniform case of either all lower case or all upper case characters [10]. After text cleaning, the next step of data preprocessing is to represent the words from the text into something that the computer can understand. One way to extract features from a given text is by utilizing the term frequency-inverse document frequency (TF-IDF) method. TF-IDF is used to represent the importance of words within a document or a corpus of documents [9] which is useful in keyword extraction, and document ranking. Another way to extract features from text data is by using word embedding. Word embedding is used to convert the text into vectors that represent the semantic meaning and relationship between

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 10, Issue 12, December 2023

words. One example of word embedding technique is Word2Vec which is used by researchers to predict words that exist in a given context [1]. With all the previous techniques mentioned, it needs to be noted that not all of them are applicable to every NLP task. Some text can be 'over-corrected' such that it loses its inherent meaning which makes it difficult or impossible for the NLP model to do analysis using that text data [5].

C. Sentiment Analysis

Sentiment analysis is a commonly used tool to extract the subjective emotion or sentiment from a user generated text. It can be used in analyzing the user experience of a given mobile application by finding out what the user really thinks of the application and its features. Researchers have tried to remove discrepancies between star rating scores and the corresponding star reviews, predict user satisfaction [1] and find out what users think are the prevalent issues in a given application [4, 5]. Sentiment analysis can be implemented both using machine learning algorithms [10] or deep learning architectures [12]. Although there are also sentiment analysis tools or models that are built on rules following a library of lexicons such as Valence Aware Dictionary and sEntiment Reasoner (VADER) which is used by researchers studying the characteristics of mobile sport apps [4].

In the case of sentiment analysis, there are further considerations to be made when doing text cleaning in order to utilize NLP methods. The first and arguably most important is the correlation between star ratings and the sentiment of the review that goes along with it. Online application platforms commonly allow the user to give a star rating alongside their written review text. This can serve as a ground truth to evaluate the effectiveness of the sentiment analysis method being used. Some researchers agree that in a 1 out of 5 star rating system, that a rating of 4 and above corresponds to a positive sentiment whereas a star rating of 2 and below corresponds to a negative sentiment [1, 3, 7]. Although a star rating of 3 can also be considered as a neutral sentiment when utilizing some sentiment analysis method [2]. Besides that, other common issues in sentiment analysis are negation handling, the imbalance between positive and negative reviews which is handled by downsampling [9] and the existence of reviews that are generated by bots which requires a filter to not cause bias in the analysis [4].

D. Topic Modeling

Finding out the different aspects of a mobile application that the users pay attention to can help when wanting to analyze the user experience of that application. These aspects may point us to what features the users are looking for in the app, and what other things may affect the user experience. Topic modeling approaches are commonly used to find these aspects and represent them as topics that are extracted from user reviews. Research in the use of topic modeling in user experience analysis varies from acting as a complement to the overall analysis user experience [3], being used in extracting the features and functionalities that exist on apps in a given category [4], for obtaining feature requests [6] and to pinpoint prevalent user complains about the application [10, 11]. Topic modeling as a base is an unsupervised learning algorithm where it does not require for it to be trained on labeled data with one of its popular implementations as the Latent Ddirichlet allocation (LDA) model used by many researchers [4, 6, 10]. Although there are simpler forms of topic modeling such as Wordcloud and n-gram models used by researchers to complement their analysis [2, 3, 12]. Despite that, there is still research done in comparing different models using labeled data [11].

In the case of topic modeling, again there are further considerations to be made when preprocessing the data before putting it through the text modeling algorithm. A research done on finding out ad-related issues in mobile applications sorts through the reviews by using regex rules and doing card sorting from the selected reviews which is then used to generate keywords that correspond to the categories that exist in ad-related issues [2]. Another research focusing on extracting feature requests from user reviews utilizes MARA which is a tool used to extract feature requests by following domain specific linguistic rules that is similar to Part-of-Speech (POS) tagging where they assign each word into a category such as 'ADJ', 'ADV', 'request' and 'existing feature' [6]. Lastly, another method used is n-gram models where sentences are broken down into word sequences of a given length. An example of a bigram is 'I eat' and an example of a trigram is 'I eat food'. These sequences are then used to statistically predict words that would appear consecutively. N-gram models are used to enhance the performance of topic modeling by some researchers [4, 10].

E. Mixed-Methods Approaches

From the above deliberation of sentiment analysis and topic modeling approaches, it is referenced that both approaches are sometimes used to build upon manual user experience analysis methods or are used in tandem to complete a certain task. One research that focuses more on a traditional approach of user experience analysis measures the quality of the application by having a group of selected raters rate several aspects of mobile applications in a given category using the an instrument called the likert scale which is then evaluated using Intra-Class Correlation (ICC) and Cronbach's Alpha to make sure that the aspects being validated are homogeneous and to validate the variance between the rating that each

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 10, Issue 12, December 2023

rater gave [3]. This paper then utilized sentiment analysis to compare the result of their rating with the actual user sentiment in the user reviews and also visualized the user reviews as a Word Cloud for both the negative and positive reviews to identify the prevalent keywords and topic of discussion for both types of reviews. Other research that combines manual and NLP based analysis methods mostly use their manual analysis as a baseline to implement the NLP method like sentiment analysis and topic modeling to categorize the rest of the data that the authors have collected [2, 9]. Moving on from the combination of manual and automated approaches, we now move our attention to the papers that combine sentiment analysis and topic modeling. The authors of both papers combine both approaches to find out the topics from a given corpus of review text and by means of sentiment analysis, they try to find out which topic was seen as the one with the worst effect on the user's user experience with on author using existing models [4] and the other proposed a system called RankMiner [5]. Other than the combining of sentiment analysis and topic modeling, some authors choose instead to focus on the data cleaning process by proposing Mining and Analyzing Reviews by Keywords (MARK) [7] and Phrase-based User Opinion Extraction (PUMA) [8], where MARK focuses on creating the basis of the data cleaning process followed by a keyword-based review analysis framework and PUMA as the follow up study where they utilizes MARK's data cleaning process to build upon a phrase-based framework to user opinion extraction.

V. CONCLUSION AND FUTURE WORK

Based on the discussion that we have above, we can conclude that the analysis of user experience is quite important shown by the amount of papers and literature that exist on the topic [1]. We also know that UX analysis can be done manually and automatically using NLP approaches. There are some papers that utilize both manual and automatic analysis [2, 9], but there are also other papers that focus more on developing systems and frameworks that utilize the full strength of NLP models [5, 8]. The development of NLP technology and the increase in user generated text online has allowed researchers and practitioners in the mobile application development field to leverage user reviews to extract user opinions and find what the users expect from mobile applications. On the other hand, many challenges have also risen with those developments which has encouraged some researchers to tackle those problems, usually in terms of localization to different languages other than English [7], and the handling of domain specific problems [6]. Researchers that use sentiment analysis use them to remove discrepancies between star ratings and actual sentiment scores extracted from user-generated text [1], compare the two measurements [3], or combine them with topic modeling to determine prevalent positive or negative topics about the applications [10]. Researchers that use topic modeling use them to find out the characteristics of applications that exist in a specific category of online applications stores, combine them with sentiment analysis, or perform tasks such as feature request extractions. With that in mind, there is still some future work that can be done to further contribute to the knowledge that exists in the use of NLP for UX analysis. Most of the papers summarize their findings in the form of insights that are given to researchers and developers of the respective applications [4]. Other than that, there are some frameworks and systems like RankMiner [5], MARK [7], and PUMA [8] that attempt to provide a tool for developers to use to leverage user reviews to determine future design considerations. Although we need to keep in mind that those tools are tailored to the English language and a specific domain or dataset with RankMiner focusing on ad performance related issues and MARK and PUMA being trained on a limited dataset. Other than that, research done on Indonesian datasets are mostly focused on calculating the performance of the NLP methods rather than the results of their processes when applied to real Indonesian mobile applications in the market. Lastly, despite the fact that there are already some frameworks and systems that provide developers with tools to analyze user reviews in terms of topic modeling, sentiment analysis, feature requests which can contribute to the development of better user experiences in mobile applications, there still aren't any tools that leverages NLP technology that attempts automatically generates things that a UX practitioner may find more useful when doing their job such as user stories and user pain points. From there with the development of image generation AI models like Dall-E [13] and Stable Diffusion [14], future work can be done to generate wireframes and designs based on the requests of the user extracted from user reviews.

REFERENCES

[1] Lee, S.H., Lee, H. and Kim, J.H. (2022) 'Enhancing the Prediction of User Satisfaction with Metaverse Service Through Machine Learning', Computers, Materials and Continua, 72(3), pp. 4984-4997. doi: https://doi.org/10.32604/cmc.2022.027943.
[2] Gao, C. et al. (2022) 'Understanding in-app advertising issues based on large scale app review analysis', Information and Software Technology, 142(106741), pp. 1-13. doi: https://doi.org/10.1016/j.infsof.2021.106741.
[3] Samad, S. et al. (2022) 'Smartphone apps for tracking food consumption and recommendations: Evaluating artificial intelligence-based functionalities, features and quality of current apps', Intelligent Systems with Applications, 15(200103), pp. 1-16. doi: https://doi.org/10.1016/j.iswa.2022.200103.
[4] Chembakottu, B., Li, H., and Khomh, F. (2023) 'A large-scale exploratory study of android sports apps in the google play store', Information and Software Technology, 164(107321), pp. 1-18. doi: https://doi.org/10.1016/j.infsof.2023.107321.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 10, Issue 12, December 2023

[5] Gao, C. et al. (2021) 'Do users care about ad's performance costs? Exploring the effects of the performance costs of in-app ads on user experience', Information and Software Technology, 132(106471), pp. 1-12. doi: https://doi.org/10.1016/j.infsof.2020.106471.
[6] Iacob, C. and Harrison, R. (2013) 'Retrieving and analyzing mobile apps feature requests from online reviews', 10th Working Conference on Mining Software Repositories (MSR), pp. 41-44. doi: https://doi.org/10.1109/MSR.2013.6624001.
[7] Phong, M.V. et al. (2015) 'Mining User Opinions in Mobile App Reviews: A Keyword-Based Approach (T)', 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 749-759. doi: https://doi.org/10.1109/ASE.2015.85.
[8] Phong, M.V. et al. (2016) 'Phrase-based extraction of user opinions in mobile app reviews', In Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering (ASE '16), pp. 726–731. https://doi.org/10.1145/2970276.2970365
[9] Ekanata, Y. and Budi, I. (2018) 'Mobile application review classification for the Indonesian language using machine learning approach', 2018 4th International Conference on Computer and Technology Applications (ICCTA), pp. 117-121, doi: https://doi.org/10.1109/CATA.2018.8398667.
[10] Permana, M.E et al. (2020) 'Sentiment Analysis and Topic Detection of Mobile Banking Application Review', 2020 Fifth International Conference on Informatics and Computing (ICIC), pp. 1-6. doi: https://doi.org/10.1109/ICIC50835.2020.9288616.
[11] Nasiri, D.F. and Budi, I. (2019) 'Aspect Category Detection on Indonesian E-commerce Mobile Application Review', 2019 International Conference on Data and Software Engineering (ICoDSE), pp. 1-6. doi: https://doi.org/10.1109/ICoDSE48700.2019.9092619.
[12] Fattahilam, A.A. et al. (2021) 'Indonesian Digital Wallet Sentiment Analysis Using CNN And LSTM Method', 2021 International Conference on Artificial Intelligence and Big Data Analytics, pp. 1-6. doi: https://doi.org/10.1109/ICAIBDA53487.2021.9689712.
[13] Ramesh, A. et al. (2021) 'Zero-shot text-to-image generation', In International Conference on Machine Learning, pp. 8821-8831. PMLR. doi: https://doi.org/10.48550/arXiv.2102.12092
[14] Rombach, R. et al. (2022) 'High-resolution image synthesis with latent diffusion models', In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695. doi: https://doi.org/10.48550/arXiv.2112.10752

AUTHOR'S BIOGRAPHY

Muhammad Zhafari Syah As an undergraduate computer science student enrolled at Universitas Gadjah Mada, I am passionate about exploring the intersection of technology and human experience. With a keen interest in both UI/UX design and artificial intelligence, I constantly strive to bridge the gap between aesthetics and functionality in digital interfaces. My academic journey has equipped me with a solid foundation in programming, algorithms, and machine learning, enabling me to approach challenges holistically

I Gede Mujiyatna was born in Jembrana, Bali, Indonesia, on April 18, 1979. He received a master's degree in Computer Sciences in 2005 from Universitas Gadjah Mada.
Currently, he is a lecturer in the Department of Computer Sciences and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia.
I Gede Mujiyatna is a member of the Association for Computing Machinery (ACM) and Indonesian Computer Electronics and Instrumentation Support Society (IndoCEISS).