A User-Centered Approach to Analyze Public Service Apps Based on Reviews

Glauco V. Pedrosa[a], John L. C. Gardenghi[b], Pollyanna C. O. Dias[c], Ludimila Felix,
Ariel V. L. Serafim, Lucas H. Horinouchi and Rejane M. da C. Figueiredo[d]
University of Brasilia (UnB), Brasilia, Brazil

Keywords: Data Mining, Machine Learning, Natural Processing Language, Human-Computer Interaction, User Experience.

Abstract: User reviews often contain complaints or suggestions which are valuable for app developers to improve user experience and satisfaction. In this paper, we introduce the Br-APPS (Brazilian Analytics of Public Posts in App Stores), an automated framework for mining opinions from user reviews of mobile apps. The purpose of Br-APPS is to assist developers by identifying the causes of negative user reviews in three levels of detail. As it is not possible to accurately estimate the number of active users of the app, Br-APPS adopts two indicators based on the number of issues reported by users and the number of app downloads. This approach allows estimating the proportion of causes of complaints, so it is possible to prioritize the issues most complained about. The performance of Br-APPS was evaluated on the gov.br app, which is the most accessed mobile application in the Brazilian government. The use of Br-APPS made it possible to quickly identify the main causes of user complaints and made it easier for developers to direct efforts to improve the app. Interviews with stakeholders of the gov.br service reported that the technique proposed is a valuable asset to ensure that government agencies can drive the creation of public services using a user-centered support technique.

1 INTRODUCTION

Mobile applications have increasingly taken a leading role in digital culture in the last years. Many governments around the world are redesigning their services to offer online public services and following this direction, mobile applications have been widely used in government domains as mobile devices are portable and facilitate access to useful services via the Internet. Advanced mobile devices, such as smartphones, iPad and tablets with apps, provide huge opportunities for public institutions to minimize the digital divide between government and citizens. The potential of mobile technologies for the public sector is manifested empirically when we observe the growing development of initiatives based on this type of device (Matos et al., 2021). Many government entities started to invest in projects to mobilize their services through mobile applications, as mobile services are ubiquitous and convenient in nature (Sharma et al., 2018).

The development process of mobile applications should take into consideration the complexities associated with their usage and exploration. Thus, developers should involve experienced users to receive feedback and to keep these constructs in focus while developing mobile applications for citizen-centric services. App markets, such as Google Play, display histograms of ratings and list text comments by users. So, these reviews contain complaints or suggestions which are valuable for app developers to improve user experience and satisfaction. However, while one could manually analyze these reviews, it is extremely tedious due to the sheer quantity of ratings and comments (Fu et al., 2013). Besides, user reviews on mobile app stores are generally shorter in length, and individual apps may have multiple releases; therefore, reviews are often specific to a particular version and vary over time (Phong et al., 2015).

One way to increase the satisfaction of citizens with public services offered via mobile application is to identify their complaints reported in the app

[a] https://orcid.org/0000-0001-5573-6830
[b] https://orcid.org/00000-0003-4443-8090
[c] https://orcid.org/0000-0003-1258-1706
[d] https://orcid.org/0000-0001-8243-7924

Pedrosa, G., Gardenghi, J., Dias, P., Felix, L., Serafim, A., Horinouchi, L. and Figueiredo, R.
A User-Centered Approach to Analyze Public Service Apps Based on Reviews.
DOI: 10.5220/0011774100003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 453-459
ISBN: 978-989-758-648-4; ISSN: 2184-4992

stores. User-centered tools that make it possible to know the failures and causes of complaints allow the public administration to direct efforts to improve its services. However, it is very common for users of public services to complain about the service offered and not about the application used to provide the service. Therefore, one of the challenges of analyzing comments from users of public services is to separate the comments that are related to the app from those related to the government. Besides that, it is also important to estimate the proportion of causes of dissatisfaction in order to prioritize utmost issues and meet users' expectations as quickly as possible.

In this paper, we introduce Br-APPS (Brazilian Analytics of Public Posts in App Stores), which is a framework for mining user opinions from user reviews of mobile apps. The technique proposed is different from other ones, mainly because it explores the data using different indicators that allow estimating the main causes of user complaints. These new indicators are valuable for prioritizing the issues most complained about by users, making it easier for developers to prioritize efforts to improve the app. Besides, Br-APPS is a multi-level framework to identify and analyze user reviews in three levels of detail: (a) the macro level shows the trending causes of complaints as well as their proportions using two metrics; (b) the meso level allows following the evolution of the main causes over time; and (c) the micro level helps to identify the set of reviews related to the causes of complaints.

The proposed approach was evaluated in a real case study with the gov.br app, which is the most accessed public app of the Brazilian government, with more than 13 million active users. The gov.br app allows Brazilian citizens to access various documents electronically, such as Driver's License, Identification number of the Brazilian citizen, Brazilian Civil Document and documents for specific audiences, such as military and civil aviation. The manual analysis of all the reviews made by the users is unfeasible due to the large volume of data; therefore, the use of Br-APPS may contribute to decision-making regarding improvements to the app.

This paper is structured as follows: Section 2 presents related works; Section 3 presents the steps of the framework developed; Section 4 shows the results obtained from an empirical experiment carried out with real data of the gov.br app; and Section 5 presents our conclusions.

2 RELATED WORKS AND MOTIVATION

Previous research showed that user feedback contains usage scenarios, bug reports and feature requests that can help app developers to accomplish software maintenance and evolution tasks. For example, the work of (Chen et al., 2014) proposed ARMiner, an approach to discover the most informative user reviews using text analysis and machine learning by filtering out non-informative reviews and topic analysis. However, (Panichella et al., 2015) argue that topic analysis techniques are useful to discover topics treated in the review texts, but unable to reveal the authors' intentions (i.e. the writers' goals) for reviews containing specific topics.

The work of (Panichella et al., 2015) presented an approach which uses Natural Language Processing, Sentiment Analysis and Text Analysis to detect and classify sentences in app user reviews that could guide and help app developers in accomplishing software maintenance and evolution tasks. The classification is performed according to a taxonomy of sentence categories deduced by analyzing reviews and development emails. (Panichella et al., 2015) empirically analyzed how the stability and fault-proneness of APIs used by some free Android apps relate to apps' lack of success. (Minelli and Lanza, 2013) proposed to combine data extracted from the app marketplace with source code to comprehend apps in depth. The work of (Chia et al., 2012) analyzed the relationship between permissions and community ratings of some apps. (Pagano and Maalej, 2013) conducted an exploratory study to analyze the user reviews from Apple Store through statistical analysis and frequent itemset mining. Also using Apple Store, (Chandy and Gu, 2012) presented a latent model to detect "bogus" user reviews. (Iacob and Harrison, 2013) developed a prototype named MARA that uses a list of linguistic rules to automatically retrieve feature requests from online user reviews.

To evaluate the users' perceptions about an app, it is necessary to obtain a reliable source of reviews that: (i) is used by a large and dispersed group of users (ii) contains information concerning the application being analyzed and (iii) provides information that discusses the problems encountered in the usability of the application. Much work has focused on detecting spam reviews (Li et al., 2011) (Mukherjee et al., 2012) (Xie et al., 2012). These works focus on detecting and removing fraudulent reviews to provide a fairer marketplace. In our work, inconsistent review detection can also help identify non-related reviews

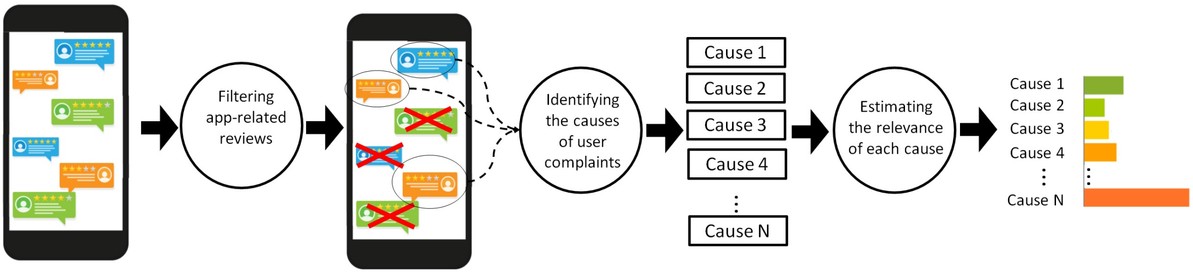
Filtering app-related reviews
Identifying the causes of user complaints
Cause 1
Cause 2
Cause 3
Cause 4
⋮
Cause N
Estimating the relevance of each cause
Cause 1
Cause 2
Cause 3
Cause 4
⋮
Cause N

Figure 1: Main steps of the proposed method to mine the user reviews on app stores.

(e.g. reviews related to the government) though our major goal is to remove non-related comments of the app in order to improve the performance of root causes discovery. Besides, it is also important to detect the inconsistencies between user comments and ratings, identify the major reasons why users dislike an app, and learn how users' complaints changed over time.

Our work also differs from the aforementioned studies because we explore the data in the app stores using other indicators to estimate the main issues reported by users. The contribution of this work is presented in the next section.

3 PROPOSED FRAMEWORK

The proposed approach, called Br-APPS, aims to mine user complaints from reviews on app stores to assist developers in improving the app. Figure 1 shows the three steps of the method proposed to identify and quantify the main causes of user complaints. In the following, we will discuss these three steps in detail.

3.1 Filtering App-Related Reviews

Reviews in mobile app stores may refer to many different issues, sometimes application-related, sometimes service-related, and sometimes government-related, especially in public service applications. This brings a variety of subjects that could not contribute to the evaluation of the quality of the mobile application itself. Distinguishing these subjects can be a tricky task. We propose a means of sorting the reviews, aiming to focus on those whose main issue pointed out is some mobile application feature.

Let $R = \{r_1, r_2, r_3, \ldots, r_n\}$ be the set of n app reviews, where each r_k consists of three pieces of information about the user feedback:

- s_k: the score of the review r_k;
- c_k: the textual comment of the review r_k;
- d_k: the date of the review r_k in YY/MM/DD format

The separation of government-related and non-government-related reviews is given by a mapping function $\phi : R \rightarrow R'$. This function is a binary trained classification-based model obtained by any machine learning classifier, where each review r_i is associated with 0 if it is non-related to the app and 1 otherwise. Thus, $r_i \in R'$ iff $r_i = 1$.

To construct function ϕ, we use a SVM classification algorithm. First, we manually build a training set by classifying each comment as application-related, service-related, or government-related. Then, we apply the identification of user complaints over the set of application-related reviews, obtaining a new set of related reviews $R' \subset R$. By filtering non-related reviews, we can improve the performance of discovering the root causes of users' negative reviews.

3.2 Identifying the Causes of User Complaints

To identify meaningful root causes, we applied the lda2vec technique (Moody, 2016), which is a combination of the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) and word2vec (Mikolov et al., 2013). We applied lda2vec to the set of related reviews R' to identify the cause-related topics.

Starting from a desired number of topics, lda2vec classifies each review in a topic by analyzing the occurrence of words that compose each topic. The words are dynamically defined during the execution of the method. lda2vec divides the set of reviews into topic clusters, and each topic cluster contains the extracted ranked top-N keywords. To meet the lda2vec input requirements, we carried out two steps:

- Step 1: One part of the input to the lda2vec model comes from the output of the LDA model, i.e., keywords, the topic-word distribution matrix and the review weights. To determine the optimal number of topics, we performed a fast manual verification using the pyLDAvis visualization method.

- Step 2: Another part of the input to the `lda2vec` model comes from the output of the word2vec model. After the process described above, we chose word vectors of the selecting keywords in the vocabulary as the other part of the input to the `lda2vec` model.

Every topic is described as a set of words related to its context. Every word contains a weight, in the form of probability values, that tells the importance of the word for the topic.

However, the absolute total number of reviews classified into each topic does not bring enough practical information. In this context, we propose two measures to identify the main topics and estimate the relevance of each cause.

3.3 Estimating the Relevance of Each Cause

We applied two measures to estimate the relevance of each problem-related topic: Problems Per Thousand Installations (PPTI) and Problems Per Thousand Daily Active Users (PPTAU). Both measures only consider the negative reviews, that is, those whose score is one or two.

Formally, PPTI is given by:

$$PPTI = \frac{\mathcal{R}}{I} \times 1000 \qquad (1)$$

where $\mathcal{R}$ is the total of daily negative reviews (eventually related to some problem) and I is the total number of daily installations of the app. The PPTI measure will give a timeless overview of all issues reported by app users. To obtain more specific temporal details of a given problem, PPTAU can be used.

PPTAU is given by:

$$PPTI = \frac{\mathcal{R}}{\mathcal{U}} \times 1000 \qquad (2)$$

where $\mathcal{R}$ is the total of daily negative reviews and $\mathcal{U}$ is the total number of daily active users in the app. On the one hand, in this way, we can check if user complaints have evolved, as have installations and active users on the app. On the other hand, we have a relative measure between complaints and installation/use of the app.

4 EMPIRICAL EVALUATION

We conducted an empirical experiment to evaluate if the approach proposed could reduce developers' efforts in real-life situations. Br-APPS was evaluated on a real case: the analysis of comments from users of the gov.br app, which is the most accessed app in the Brazilian government. It has been downloaded by more than 22 million users, which continues at an increasing rate. Between 2021 and 2022, many of its users reported difficulties due to problems in the app or even low digital literacy. User complaints are recurrent and abundant in the app stores, which makes it hard to perform a manual analysis and practically impossible to transform the user's issues into information for decision-making regarding improvements to the app.

Our goal is to evaluate if the Br-APPS can assist stakeholders of gov.br to improve the app complying with the users' issues reported in their reviews on the app stores.

4.1 Data Collection and Processing

We used an open-source crawler to continuously obtain the most recent reviews of the gov.br app. We crawled 44,500 reviews on Google Play from November 4, 2020 to August 30, 2022. Each review crawled contains a title, a long text description of its content, the time of creation, the reviewer ID, and its associated rating.

Before identifying the causes of user complaints, we applied standard filters in the set of reviews:

- *Tokenization*:First, all reviews were tokenized. This is a process in which a stream of text is broken into words, phrases or other meaningful units. This process was done with a simple python NLTK library. Punctuations were removed from the sentences, then the text stream was split up by whitespace characters. Other non-word and non-number characters were also removed from the tokenized dataset.
- *Stopwords removal*. Stopwords are common words that do not have semantic relevance to the context analyzed. These are usually prepositions and articles.
- *Lemmatization*. This consists of transforming words into a basictheir base form, called lemma. For instance, words *am, is, are* would bewere reduced to the lemma *be*.

Figure 2 shows the overview of the app reviews collected in the period considered, as well as the increase in the total number of evaluations over

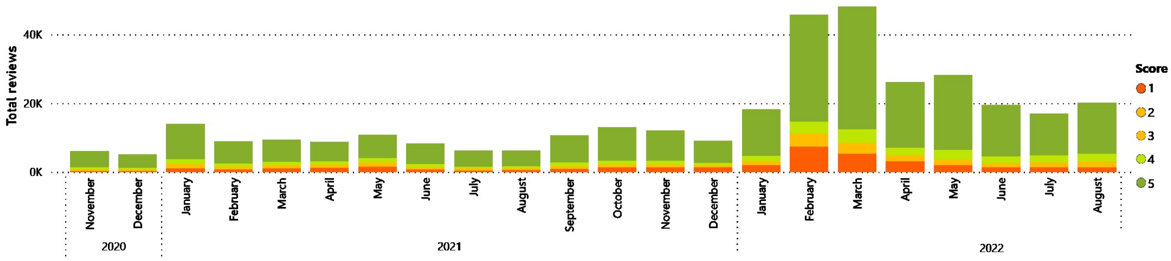

Figure 2: Total monthly reviews per score.

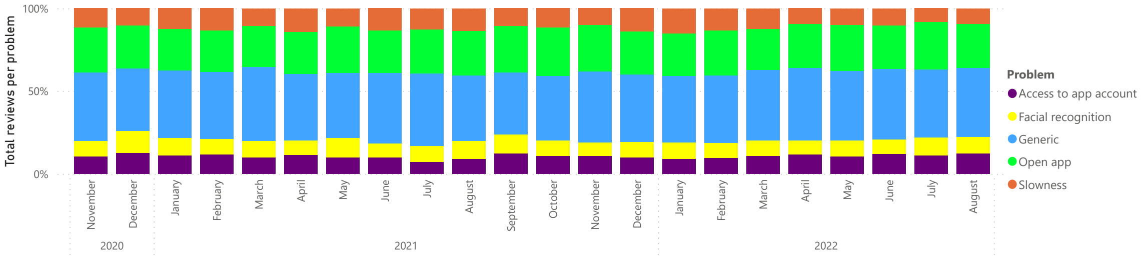

Figure 3: Total monthly reviews per problem.

time. Although negative reviews are the minority of the evaluations, understanding their root is very important for the continuous improvement of the service offered. In 2020-2021, the app had an average of 3 million users per month. Today, it has an average of 17 million users. Despite a large number of negative comments, the store rating is stable.

4.2 Obtained Results

From the data provided by Br-APPS, we built 4 graphs to visualize the results obtained. These graphs are represented in Figures 3, 4, 5 and 6. In the following, we will discuss the information extracted from these graphs.

We identified problems in five main aspects reported by users of the gov.br app: (i) initialization; (ii) slowness; (iii) login; (iv) facial recognition and (iv) generic causes. Figure 3 shows the distribution of these problems per month at the meso level. In 2020, a version of the app was released with face recognition authentication. Many users installed the app, but faced trouble in the account creation process, which lead them to leave negative comments about the app. Subsequently, the app was completely overhauled to provide a better user experience. The number of negative comments began to decrease, and the comments stabilized, as shown in Figures 4 and 5.

At the micro level, Figures 4 and 5 show the PPTI and PPTDAU measures per month in the period considered. Both measures decreased in this period even with the increased number of reviews. This shows that the relative satisfaction of the app users has grown over time. Analyzing the results, we were also able to identify some instabilities in the app, which stakeholders attributed to an unusual number of people accessing the app. In February 2022, the period for filing the Brazilian Tax Income Declaration began. Citizens can use a pre-filled declaration when they have a gold or silver gov.br account. The requirement for gold and silver accounts enhanced the number of active users of the platform and app installs. Also since February 2022, many users attempted to access the System on Receivable Values, which also requires a gold or silver gov.br account. This is a service from the Brazilian Central Bank where one can check if they have some forgotten or unexpected money in a bank or other financial institution and, if so, redeem that amount.

Finally, Figure 6 illustrates the distribution of negative reviews at the macro level in five aspects.

4.3 Perceptual Evaluation

We interviewed eight stakeholders of the gov.br app to evaluate their perception of Br-APPS. The interviews were carried out using a questionnaire with the following agree/disagree survey questions:

- Q1: Br-APPS was able to detect the main factors of user dissatisfaction with the app
- Q2: Br-APPS made it possible to identify details of the causes of user dissatisfaction with the app
- Q3: Br-APPS helped the team responsible for the

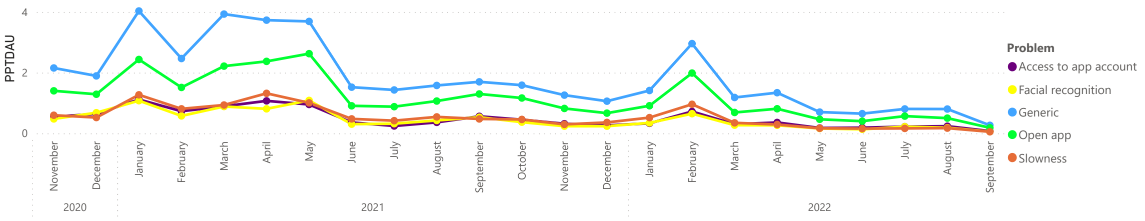

Figure 4: Total number of Problems Per Thousand Active Users (PPTAU) per month.

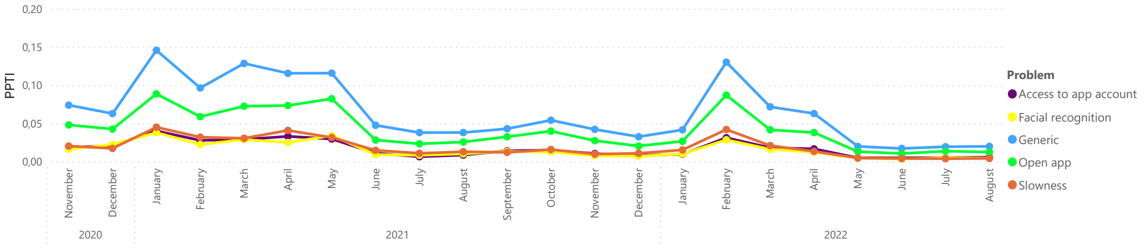

Figure 5: Total number of Problems Per Thousand Installations (PPTI) per month.

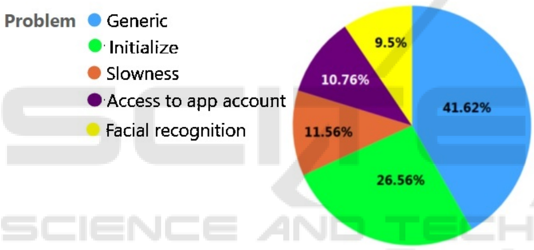

Figure 6: Total reviews per problem.

app to direct efforts to improve the app

- Q4: Br-APPS was able to provide input on reasons for installing and uninstalling the app
- Q5: Br-APPS was able to present the most and least accessed features of the app

Each interviewee should answer each question using the Likert-5 scale: totally agree (5), partially agree (4), neutral (3), partially disagree (2) and totally disagree (1).

Table 1 shows the questions and the average score obtained with the responses. According to the results, we observed that using Br-APPS can contribute to assisting the developers to map the issues reported by users while providing a reliable way to search for the most relevant ones. In their opinion, Br-APPS managed to save them time and effort in discovering and understanding users' opinions but failed to discover more detailed semantic meanings of user' reviews. Unfortunately, this drawback is not related to Br-APPS functionality, as it is not always possible to get details of user complaints due to the short texts they provide in the app stores.

Table 1: Average score obtained for each question.

Question	Average Score
Q1	4.250
Q2	3.875
Q3	4.125
Q4	4.000
Q5	3.625

5 CONCLUSIONS

In this paper, we presented a text mining based approach to identify and estimate the main causes of user complaints from app store reviews. The method proposed, called Br-APPS, allows to evaluate the main topics of user complaints in three levels of detail and, thus, help developers to improve usability issues and user experience design. Br-APPS can offer important insights that benefit end users, developers and potentially the entire mobile app ecosystem. More specifically, we provide data that allow people to easily absorb the information contained in a large set of text reviews and numerical ratings by offering multiple forms of summarization.

The main contribution of Br-APPS is two indicators based on the number of daily installations of the app and daily active users in the app. Thus, it is possible to have a timeless overview of all the issues reported by the app user while checking if user complaints have evolved. These two new metrics play a crucial role in providing a dynamic view of how users' opinions evolve, therefore discovering event-driven trends and life spans of different versions.

The findings obtained by Br-APPS allowed its stakeholders to direct their development efforts towards the main issues reported by users. Br-APPS was able to save them time and effort in discovering and understanding users' opinions so that governments could harness the potential of digital technology and data to improve outcomes for all.

ACKNOWLEDGMENTS

This work has been supported by Ministério da Economia (ME) - Secretaria de Governo Digital (SGD) Transformação Digital de Serviços Públicos do Governo Brasileiro.

REFERENCES

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Chandy, R. and Gu, H. (2012). Identifying spam in the IOS app store. WebQuality '12, page 56–59, New York, NY, USA. Association for Computing Machinery.

Chen, N., Lin, J., Hoi, S. C. H., Xiao, X., and Zhang, B. (2014). Ar-miner: Mining informative reviews for developers from mobile app marketplace. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE 2014, page 767–778, New York, NY, USA. Association for Computing Machinery.

Chia, P. H., Yamamoto, Y., and Asokan, N. (2012). Is this app safe? a large scale study on application permissions and risk signals. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 311–320, New York, NY, USA. Association for Computing Machinery.

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. KDD '13, New York, NY, USA. Association for Computing Machinery.

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.

Li, F., Huang, M., Yang, Y., and Zhu, X. (2011). Learning to identify review spam. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, page 2488–2493. AAAI Press.

Matos, E., B. B. Lanza, B., and D. Lara, R. (2021). Mobile government in states: Exploratory research on the development of mobile apps by the brazilian subnational government. In *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, DG.O'21, page 351–362, New York, NY, USA. Association for Computing Machinery.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781v3.

Minelli, R. and Lanza, M. (2013). Software analytics for mobile applications–insights & lessons learned. In *2013 17th European Conference on Software Maintenance and Reengineering*, pages 144–153.

Moody, C. E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv:1605.02019.

Mukherjee, A., Liu, B., and Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 191–200, New York, NY, USA. Association for Computing Machinery.

Pagano, D. and Maalej, W. (2013). User feedback in the appstore: An empirical study. In *2013 21st IEEE International Requirements Engineering Conference (RE)*, pages 125–134.

Panichella, S., Di Sorbo, A., Guzman, E., Visaggio, C. A., Canfora, G., and Gall, H. C. (2015). How can i improve my app? classifying user reviews for software maintenance and evolution. In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 281–290.

Phong, M. V., Nguyen, T. T., Pham, H. V., and Nguyen, T. T. (2015). Mining user opinions in mobile app reviews: A keyword-based approach (t). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 749–759.

Sharma, S. K., Al-Badi, A., Rana, N. P., and Al-Azizi, L. (2018). Mobile applications in government services (mg-app) from user's perspectives: A predictive modelling approach. *Government Information Quarterly*, 35(4):557–568.

Xie, S., Wang, G., Lin, S., and Yu, P. S. (2012). Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 823–831, New York, NY, USA. Association for Computing Machinery.