Machine learning models identify apps that will likely violate Google Play store guidelines

October 13 2022

Credit: Pixabay/CC0 Public Domain

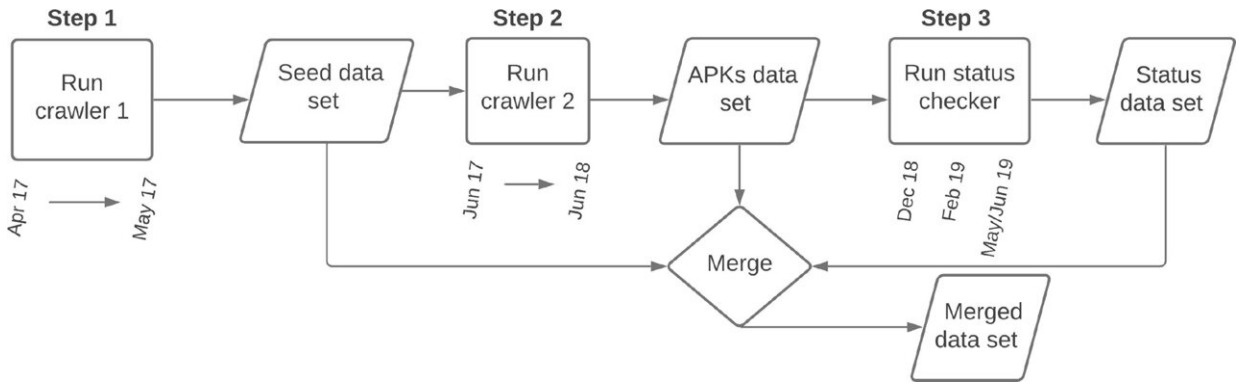
A considerable percentage of new apps in the Google App store are removed for violating the store's guidelines. This is inconvenient for the

users of these apps, who may lose their in-app data. Computer scientists from the University of Groningen have devised two machine learning models that can predict the chances of a new app being removed, both before and after uploading it to the app store. These models can help both developers and users. The details of this project are described in a paper that was published in the journal *Systems and Soft Computing* on Sept. 29.

The Google Play store has set rules and requirements that developers must adhere to. After being submitted, apps are immediately uploaded to the store, but it takes Google some time to vet them before they remove apps that are found to violate the guidelines. Developers whose apps have been removed more than once, may face a ban from the store.

"My research interest lies in digital privacy and security issues," says Fadi Mohsen, assistant professor at the Information Systems Group of the Bernoulli Institute for Mathematics, Computer Science, and Artificial Intelligence, University of Groningen. Given the consequences of app removal for both developers and users, he wanted to create a system that would be able to predict whether new apps will be removed or not.

"There have already been attempts to do this, but these typically focus on specific types of apps that were removed for specific reasons, for example because they contained malware," Mohsen explains. "We wanted to develop a general model that predicts the chances of an app being removed, regardless of the type of app or the reason for removal." Furthermore, previous attempts focused solely on users, while Mohsen also wants to assist developers who just fell foul of the guidelines by accident.



A high-level overview of the data collection workflow. Credit: *Systems and Soft Computing* (2022). DOI: 10.1016/j.sasc.2022.200045

The first step was to gather a large data set from apps that were removed and of apps that were not removed: "We collected metadata, including the descriptions provided by the developers to the store, from roughly two million apps. After that, we downloaded the source code of half of these apps."

Subsequently, Mohsen and his colleagues tracked the status of these apps in the store for six months to see which apps were removed. "In our selection this was the case for 56 percent of them." It took them 26 months to finalize the data set that was used to generate the machine learning models.

The algorithm they used is called Extreme Gradient Boosting. "It is the best machine learning algorithm for these kinds of problems," explains Mohsen. The algorithm was used to create two predictive models: one for developers and one for users. The model for users was determined by 47 features, and in a test data set it predicted the removal of a given app with 79.2 percent accuracy. As some of these features, like ratings in the app store, are not available before submitting the app to the store, the

developer model was based on only 37 features, and its accuracy was slightly lower as a result: 76.2 percent.

"We can now predict the future of an app with reasonable accuracy," says Mohsen. The next step is to develop an interface with which developers and users can assess apps on their risk of removal. "This is valuable for developers, as they could be banned from the Google App store if they violate the guidelines repeatedly," says Mohsen, 'but also for users, as they generate data with their apps, which they will lose if the app is suddenly withdrawn."

Other researchers will also benefit from this research. "The rich data set we have generated for our paper has been made publicly available through the Dutch repository Dataverse.nl," says Mohsen. This means that anyone can try to improve on the results obtained by Mohsen and his colleagues. "We are looking forward to the competition, to find out if they can beat us. That would further increase the benefit for users and developers."

More information: Fadi Mohsen et al, Early detection of violating Mobile Apps: A data-driven predictive model approach, *Systems and Soft Computing* (2022). DOI: 10.1016/j.sasc.2022.200045

Dataset: dataverse.nl/dataset.xhtml?persistentId=DOI: 10.34894/H0YJFT

Provided by University of Groningen