Windows app store analytics: A case study on the windows app store dataset

Govinda Roy, MD. Rasidul Haq Shuvo, Fatin Ishraq, Zahiduzzaman Pranta, MD. Shamsur Rahim
Department of Computer Science, American International University – Bangladesh
Govindaroy.ofc94@gmail.com

***Abstract*— Installation of fraud app refers to the scenario when the users want to install an app from store according to their choice of use but after installation they can see that the app is not as useful as their expectation. Moreover, this kind of app can be full of viruses or malwares which are very harmful for their devices and can steal the information of the users for bad reasons. Since there are huge number of apps in different play stores, these cannot be stopped manually. The users need a reliable platform where they can know which app is more helpful. In this paper it is presenting that by using a web scrapper and huge amount of data from windows store how fraud app can be detected. Doing judgement by watching other users' reviews is not time saving. Sometimes some apps lose their track and cannot give service like previous. It can be also found that out by doing scrapping to help the users. The authors believe that this is the way how the users can keep their information secure and use virus free devices. Furthermore, this paper identifies research gaps in the existing works and provides future research directions.**

***Index Terms*— Windows app store, windows app store dataset.**

I. INTRODUCTION

In the modern era smart devices are part and parcel of our daily life and we use them to make our life easy. In smart devices we keep the passwords of our social accounts and also the passwords of our bank. To do different things and for various purposes we install different types of apps from app store. Sometimes, after installation we see that the installed app cannot fulfil the desires requirement and also often they are full of unwanted advertisements. These can be a trick for many hackers to steal our useful information for really bad purposes. If the worthiness of the app is satisfactory then we keep the app in our device and keep using it until we do not have use of it anymore.

Popular apps can give us satisfactory results but it can also contain viruses or malwares which can be very costly for us. Often the user cannot do anything but uninstalling the app and move in wile the apps are not performing well bit malwares remain in the smart devices. When we keep these kind of apps in our mobile, day by day they steal information from our mobile or smart devices. It is a way to clarify before downloading by reading the comments about desired apps and watching the rating of it but this is very much time consuming and sometimes wrong and pad comments can confuse us. What if a dataset of those apps can help us? If this dataset is used for a system which can show us which apps are fraud or rogue then it would be very much easier for us not waste our time rather not face those consequences.

We have analyzed the app stores there are. In these app stores we saw that, Google Play Store and Apple Play Store do not keep the apps in their store which have below three ratings. There are already some systems generated in their store to control this issue. But in the Windows App Store, we can find the apps of all criteria and ratings. Here we can find more data which contains fraudulence. By building a web scrapper we collected the data from Windows Store. Here we are considering the following attributes for the dataset:

- App Category
- App Title
- App Publisher
- App Rating
- App Price
- Comment Rating
- Comment Date
- Reviewer Name
- Platform
- Comment Title
- Full Comment
- Usefulness

The following sections discussed in this paper are structured as follows: background study and related works in Section II, data collection method in Section III, description of dataset in section IV, data overview in Section V and future research direction and conclusions in Section VI.

II. BACKGROUND AND RELATED WORKS

The very first thing is needed to understand is, why we are collecting information from the Windows app store? The answer is very simple. We believe that, windows app store has rogue apps. And the apps are collecting data from the users without their knowledge. The AR-miner [1] performs comprehensive analytics from raw user reviews. MSR [2] has the same concept as well. A systematic literature review on mining online reviews from mobile apps stores [3], has identified proposed solutions for mining online opinions in app store. It has been found that, many apps increase their version-to-version rating, while the store-rating of an app is

resilient to fluctuations once an app has gathered a substantial number of raters [4]. Wiscom [5] which is a system that can analyze user ratings and comments in an app store at three different levels of detail.

III. DATA COLLECTION METHOD

The data is collected from the Windows App store website by using a web scrapper. The data were crawled from the raw-HTML document website. The data collection procedure is shown in Figure 1.

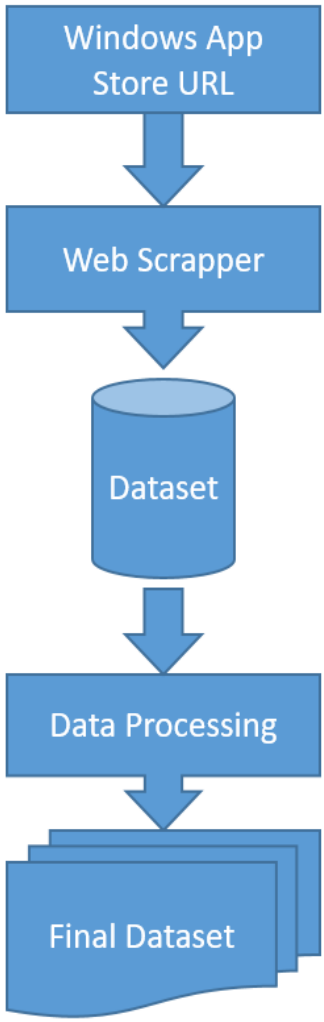

Figure 1: Data Collection Procedure

A. Windows App Store URL
The URL that is being used for scrapping the data is at https://www.microsoft.com/en-us/store/apps/windows?icid=CatNavWindowsApps. There are total 25 category in the app store. Each category contain at least 111 apps.

B. Web Scrapper
We have used a web scrapper for extracting the HTML page of Windows App store. To create the crawler, we have used Selenium WebDriver, an open source tool and programmed it with C#. The scrapper is lightweight automation tool. In each requested page, the scrapper can pass through DOM [6].

C. Data Extraction
The information were extracted from the HTML document. The Selenium WebDriver tool crawls the information from the particular elements of the DOM. We also have used, Regular Expression (RE) in order to find the part of that particular element.

D. Local Storage

E. The dataset format is saved in Comma Separated Value (CSV) format which is stored in local storage. The reason for choosing this format for ease of reading the dataset and the attributes' value each instances.

F. Data Processing
The raw dataset that was collected may have duplicate, errors, inaccurate or missing values. These inconsistencies must be removed for accurate and efficient result. We processed the data by using OpenRefine [7].

G. Final Dataset
Until now we have were able to crawl 15355 of instances. More instances will be contributed to the dataset. Due to the alpha beta version of windows app store, the time for crawling the data was very much time consuming. So, we could not crawl each and every category from the URL.

IV. DESCRIPTION OF DATASET

After crawling the dataset of Windows App store, the dataset contains the following attributes:

1. Category: The name of the category of the app.
2. Title: The title of the app.
3. Publisher: Company that developed the app.
4. Rating: The average rating of the app.
5. Price: Is the app free or not.
6. Comment Rating: The rating of each user.
7. Comment Data: The date when was the comment posted.
8. Reviewer Name: Name of the reviewer who posted the comment.
9. Platform: From which platform the user posted the comment.
10. Comment Title: The title of the comment.
11. Full Comment: The comment that describes the user's review.
12. Usefulness: To how many people the comment given by the user was useful to.

Here in Figure 2, a small fragment of xml schema of the dataset is given.

```
<?xml version="1.0" encoding="UTF-8"?>
<root>
  <row>
    <categoryName>Personal finance</categoryName>
    <appTitle>Spending Tracker ™</appTitle>
    <publisher>MH Riley Ltd</publisher>
    <rating>4.4</rating>
    <price>Free+</price>
    <commentRating>5</commentRating>
    <commentDate>3/6/2016</commentDate>
    <reviewerName>john aibert</reviewerName>
    <platform>PC</platform>
    <titleComment>Helpful</titleComment>
    <fullComment>Helps me manage my future expenses,gives me a smooth over view of my financial sit.</fullCom
    <usefulness>3 out of 5 people found this helpful.</usefulness>
  </row>
```

Figure 2: XML Schema of Dataset

V. DATA OVERVIEW

Here, we are going to provide some insights of the dataset. The Windows App store dataset contains 21544 instances (until now). The dataset contains 5 categories. For now, we have managed to collect Health & fitness, Kids & family, Medical, Personal Finance, Social categories from the app store.

Here in Figure 3, we have counted the number of comments in the category according to our dataset with a Total number of Comments of 15355.

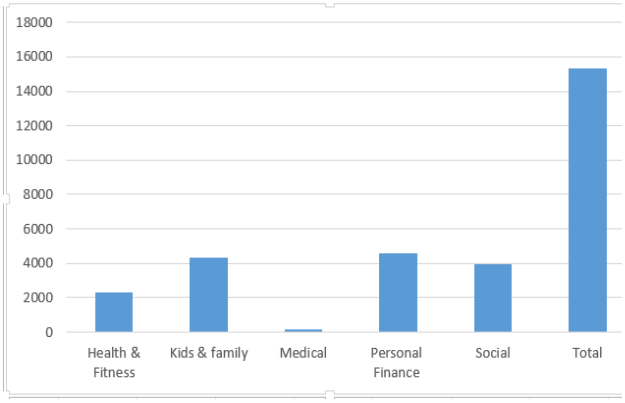

Figure 3: Number of Comments on each Category

VI. LIMITATIONS

The scrapper could not crawl the user comments which had long length. Due to different versions of windows apps store, we had to do alpha beta testing with the website. During crawling the website, the website changes frequently. If the change happens, the scrapper could not crawl the information. Thus exceptions occur during the running the code. Which is the very first reason for taking almost 7+ hours crawling the data from just one category. In some cases, the apps which had too many reviews, the scrapper could not crawl all of them. Because one refresh changed the whole interface during crawling the reviews. It took more on others. The process is very time consuming. And we took the data from each category one by one. From the dataset, we are only taking the reviews which are English. Reviews of other languages were neglected.

VII. FUTURE RESEARCH DIRECTIONS AND CONCLUSIONS

The dataset can be analyzed for predicting fraud apps and rogue apps or malwares. Also, by using the dataset a ranking system can also be generated. Not only that, through analysis, the dataset can provide us which app is good based on the user comments. Or which app has stopped working. So, various types of queries can be done from this dataset. So, our dataset can be implemented to:

- Develop predictive accuracy on identifying fraud apps.
- Making website like Appbot for ratings and reviews.

The data gathered from the windows app store is very important for further research on Windows app store.

References

[1] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang, "AR-miner: mining informative reviews for developers from mobile app marketplace," *Proc. 36th Int. Conf. Softw. Eng. - ICSE 2014*, pp. 767–778, 2014.

[2] M. Harman, Y. Jia, and Y. Zhang, "App store mining and analysis: MSR for app stores," *IEEE Int. Work. Conf. Min. Softw. Repos.*, pp. 108–111, 2012.

[3] N. Genc-Nayebi and A. Abran, "A systematic literature review: Opinion mining studies from mobile app store user reviews," *J. Syst. Softw.*, vol. 125, no. November, pp. 207–219, 2017.

[4] I. J. Mojica Ruiz, M. Nagappan, B. Adams, T. Berger, S. Dienst, and A. E. Hassan, "Examining the Rating System Used in Mobile-App Stores," *IEEE Softw.*, vol. 33, no. 6, pp. 86–92, 2016.

[5] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh, "Why people hate your app," *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '13*, p. 1276, 2013.

[6] H. Borland and J. Kilmer, "Document Object Model ™ (DOM)," *I Can*, 2001.

[7] D. Huynh, K. Points, T. Interactions, C. Editing, and S. Editing, "Google Refine," *Text*, pp. 1–17, 2011.