Detection of Rental Listing Scams Using Machine Learning

Sargam Ghorela
Student ID: 1300455
sghorela@nyit.edu

Saurav Ganguly
Student ID: 1288841
sgangu01@nyit.edu

I. Problem Statement

In the present world, several rental platforms provide accommodation options to their users. Rental platforms are gaining popularity nowadays and constitute a significant part of the individuals planning to book a stay. Booking a visit with any property is a massive task, and booking through a rental platform makes it feasible and problematic simultaneously. Many factors are involved in booking a stay: amenities, neighbourhood, location, reviews, description, the property itself, and its price. These online rental platforms give us much information about every comfort and image to plan better, but sometimes this information needs to be more accurate. Several scams and frauds occurred on many rental platforms, leaving the customer unsatisfied and in a financial crunch for a few days or over a month. Several listings on the platform don't exist legally, and others are "not as described." The rental platforms are trusted upon their user review and host review ratings, which assures customers that they are in the right place, but these trusted factors are also rigged. There are plenty of reviewers and hosts, which makes the platforms more untrustworthy. One of the renowned platforms, Airbnb, also faces these loopholes in its system. According to 127,183 Airbnb customer complaints analysed by Asher & Lyric, it is specified that more than 40 percent of the listings fall under Multiple listing scams. Apart from these, 25 percent are described as "Not as Described"[4]. These scam numbers are concerning and need an organized detection system

II. Review Of Previous Work

Customer satisfaction is an important goal for rental platforms, So these scams are a major loophole that should be detected and removed. Unfortunately, no useful detection techniques have been developed to recognise these scams other than reporting a significant issue and getting the listing removed or, in some cases, the host. One of the data scientists' research is based on the network created by connected hosts which publish listings[3]. The article calls it systemized listings which means that some concerned accounts with reviews from a small group of reviewers host listings. As per the wired articles, this is run by some organizations to pump money through systemized listings[2]. According to this analogy and data representation with the help of networks, there is proof that the published article somehow signifies this as a major problem and an important feature to establish a fake listing.

III. Project objective and methodology

In rental platforms, trustworthiness is more important for the customer in every aspect, from comfort to financial necessities. To fulfil these particular causes, a way to provide users with genuine listings to choose from is the primary concern. The targeted platform, according to the project, i.e., Airbnb, has major legal compliance issues due to the region-specific limitations and some organizations running a systematic listing approach by creating duplicate listings on the platform. So the first approach for the use case multiple listing would be to train the model to mark the duplicate listings under the suspicion category and ask for more details to certitude the property. According to data provided by Airbnb for quality and assurance purposes insideairbnb we have data related to listing id, which is divided and stored under reviews.csv ,neighbourhood.csv, and listing.csv[5]. The data provided contains the artifacts related to the property description, amenities, neighbourhood, host acceptance rate, host data, a number listed provided under hosts, reviewer's artifacts, images of the property, etc., for every listing they have for a certain year. This data is to be categorized with the help of certain features like images, addresses, descriptions, and so on using vectors and variables, which can be numerical, categorical, and text[8]. The data categorized will be cleaned by removing duplicates, irrelevancy, and unwanted punctuations, followed by tokenization using the Natural Language toolkit, spaCy and regular expressions. Stemming is also needed for some data artifacts before feeding them to the machine-learning model. After the data is fed, indexed filtering will be performed to max out the top features, and a stack ranking will be developed using a deep neural network that runs a matching algorithm for the elements. Deep learning architecture Autoencoders[1] would check the image similarity for the above multiple listing[7]. This is a feedforward neural network used for compression or dimensionality reduction. The whole process will lead to an inference to apply classification to calculate a similarity score to mark the listing under the suspicion category[6]. The threshold for this similarity check will be defined in the later stages of the project[7]. The next user case is to train the model to point out the properties which are "not as described" or fake for this, we our pointing out the host and reviewer relations by defining the different sets that include where reviewers have multiple reviews for the same host, hosts with multiple listings and reviewers who reviewed multiple hosts. This graphical representation would be merged with the host acceptance rate, and reviews made description categorization to check for the possible spammer host and reviewers, which could be done using a graph neural network to create an inference and for further regression. This will also mark the suspicious host and reviewer, which needs to be certified by the platform on their authenticity[6].

IV. Resourses Used

We are considering using Python & R language for this project to write the required code. For Specific multiple listing problems using machine learning to filter out the text,

we will be using Data Cleaning to remove unwanted characters, spaces, or HTML tags, remove duplicates, irrelevant data that can interfere with the analysis, Tokenization, Stop Words Removal, and Stemming/Lemmatization using NLTK (Natural Language Toolkit), spaCy libraries or regular expression. For the model, we will be using Indexed filtering to filter out the query from the pool quickly. We will be using Deep Neural Networks, Feed Forward Neural Networks, Graph Neural Networks, Classification and Regression models.

V. Contribution To Society

Rental platforms such as Airbnb gained massive popularity among people due to their facilities and features and are a boon to society. However, some scammers use the loopholes to make money unethically. Our solution will help remove the spammed listings from the rental platforms, which can provide a very trustworthy and authentic experience to the customers. This will help the customers organize and plan their stay better without falling under any scammer's trap and financial holds they currently face. Our machine learning model will also identify fake reviews and help the rental platforms remove such reviews so that customers don't get misguided by paid or fake reviews, which is a huge problem nowadays.

References

[1] Arden Dertat, "Applied Deep Learning-Part3:Autoencoders," retrieved on 27th Feb 2023, https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

[2] James Temperton, " I stumbled across a huge Airbnb scam that's taking over London.," retrieved on 27th Feb 2023, https://www.wired.co.uk/article/airbnb-scam-london

[3] Julia McAleenan, " Identifying potential scam listings on Airbnb," retrieved on 27th Feb 2023 , https://towardsdatascience.com/identifying-potential-scam-listings-on-airbnb-e9aed41611e5

[4] Asher Fugersson, "127183 Airbnb Guest complaints Expose Scams, Safty Concerns, Infestations & more", retrieved on 28th Feb 2023 , https://www.asherfergusson.com/airbnb/

[5] Inside Airbnb, " Get the data", retrieved from 27th Feb 2023, http://insideairbnb.com/get-the-data/

[6] Ameya," Applciaiton of ML:Building Indices?", retrieved on 28th Feb 2023,https://medium.com/coinmonks/building-efficient-learned-indices-using-machine-learning-96890c0fa948

[7] David Loshin,"Similarity Score", retrieved on 28th Feb 2023, https://www.sciencedirect.com/topics/computer-science/similarity-score

[8] Formplus,"Categorical & Numerical Data: 15 Key Differences & Similarities", retrieved on 1st March 2023, https://www.formpl.us/blog/categorical-numerical-data