Text Mining
Project Handout
2023

Predicting Airbnb Unlisting

1. Project Objectives

The goal of this project is to use Natural Language Processing (NLP) models to predict whether a property listed on Airbnb will be unlisted in the next quarter[1]. For this, you will resort to real Airbnb property descriptions, Airbnb host descriptions, and comments from previous guests. In summary, with the NLP techniques you will learn during the Text Mining course, you must implement an NLP classification model able to predict, for each property, if it was unlisted (1) or is still listed (0).

The project should be developed using Python and libraries such as NLTK, Scikit Learn, Keras or PyTorch. Also, the project can be solved in various ways, which means there is no exact correct solution. And, of course, groups should not use code from each other!

2. Group Rules

The project is to be developed individually or in groups of two (2) to four (4) students.

3. Corpora

The data is divided in following sets:

- **Train** (train.xlsx) (12,496 lines): Contains the Airbnb and host descriptions ("description" and "host_about" columns), as well as the information regarding the property listing status ("unlisted" column). A property is considered unlisted (1) if it got removed from the quarterly Airbnb list and it is considered listed (1) if it remains on that same list.
- **Train Reviews** (train_reviews.xlsx) (72,1402): This file has all the guests' comments made to each Airbnb property. Note that there can be more than one comment per property, **not all properties have comments, and comments can appear in many languages!**
- **Test** (test.xlsx) (1,389 lines): The structure of this dataset is the same as the train set, except that it does not contain the "unlisted" column. The teaching team is keeping this information secret! **You are expected to provide the predicted status (0 or 1) for each Airbnb in this set. Once the projects are delivered, we will compare your predictions with the actual (true) labels**.
- **Test Reviews** (test_reviews.xlsx) (80,877): The structure of this dataset is the same as the train reviews set, but the comments correspond to the properties present on the test set.

[1] Original data retrieved from: http://insideairbnb.com/

Note that you are not required to use all the textual fields from the provided corpora as input for your model. For example, your best solution may just use the “description” column as input.

4. Solution Requirements and Evaluation Criteria:

Your solution should present the following points:

1. **Data Exploration** (**1.5 points**): Here you should analyze the corpora and provide some conclusions and visual information (bar charts, word clouds, etc.) that contextualize the data.
2. **Data Preprocessing** (**2 points**): You must apply a method to split your training corpus into train/validation sets to evaluate the performance of your model (you can also resort to K-Fold cross validation, or other methods). Moreover, you must correctly implement and experiment at least four (4) of the data preprocessing techniques shown in class (stop words, regular expressions, lemmatization, stemming, etc.).
3. **Feature Engineering** (**5 points**): You must correctly implement and experiment with two (2) of the feature engineering techniques seen in class (TF-IDF, GloVe embeddings, etc.).
4. **Classification Models** (**4.5 points**): You must correctly implement and test three (3) of the classification algorithms seen in class (KNN, LR, MLP, LSTM, etc.).
5. **Evaluation** (**1.5 points**): You must evaluate your models resorting, at least, to Recall, Precision, Accuracy and F1-Score.

Moreover, the development of extra work (more techniques than the minimum required in the previous points and/or techniques not shown in class) is highly recommended and will account for a maximum of **4.5 points** divided as follows:

1. **Data Preprocessing – 0.25 points** for each extra method (unseen in class) used (maximum of 2 extra methods).
2. **Feature Engineering – 1 point** for each extra method using Transformed-based embeddings (maximum of 2 extra methods)
3. **Classification Models – 1 point** for each extra model using Transformers or more advanced models (maximum of 2 extra methods).

5. Delivery Guide

In terms of the solutions developed, you must deliver:

1. One .pynb file (notebook), named NLP_XX (XX stands for the group number), containing the techniques you experimented and, by the end of the notebook, your ready-to-run final solution.
2. A .csv file, named “Predictions_XX”, with the Ids of the test set and your predicted labels for the test set.

Additionally, you **must submit a PDF report** named “Report_XX”, documenting your work, with the following structure (other structures are also accepted):

1. **Data Exploration** – data presentation and explanation of the main finding from the exploratory analysis (accounts for **50%** of criteria **4.1**).
2. **Data Preprocessing** – explanation of the different preprocessing methods developed (accounts for **25%** of criteria **4.**Error! Reference source not found.).
3. **Feature Engineering** – description of the methods implemented (accounts for **25%** of criteria **4.3**)
4. **Classification Models** – description of the models implemented (accounts for **25%** of criteria **4.4**)
5. **Evaluation and Results** – description of the performance of the models and main conclusions (accounts for **50%** of criteria **4.5**)

The PDF report should have a maximum of 10 pages describing the previous points. Exceeding this number will incur a **0.5-point penalty** for each extra page.

Any **extra work** developed **must be clearly defined as such in the PDF report**, or else it will not be considered for evaluation as extra work! You should add, in each section, a paragraph pointing out the extra work developed.

All files should be saved in a folder named “Group_XX”. This folder (zip it if you need) must be submitted through the project submission section in Moodle, until **23h:59 of the 18th of June (Sunday)**.

Failure to deliver on time will incur a **1.0-point penalty** for each half-day late.

Failure to comply with the delivery guide will meet with a **0.5-point penalty**.

Final Notice:

We will compare your predictions with the actual Label from the test set (“test.csv”).

The three (3) groups with the highest performance will receive points as follows:

- **1 point** for the group with the best model
- **0.5 points** for the group with the 2nd best model
- **0.25 points** for the group with the 3rd best model

Students may be randomly selected for an **oral defense** to access their knowledge.

Good luck with your project!