NOVA IMS Information Management School

MDSAA

Master Degree Program in

Data Science and Advanced Analytics

Text Mining

Predicting Airbnb Unlisting

Andrei, Macovei,number: 20221358

Group 26

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

June, 2023

INDEX

INTRODUCTION

The way people travel and find accommodation has completely changed thanks to Airbnb, a well-known web site for accommodations and holiday rentals. Since its founding in 2008, Airbnb has provided a platform that links hosts with available rooms with visitors looking for distinctive and customized accommodations. Airbnb is growing to be a popular choice for tourists looking for a more unique and genuine stay because to its wide range of property types, which includes cozy apartments, luxurious villas, ecentric treehouses, and attractive cottages.

In terms of the project at hand, my main goal is to use Natural Language Processing (NLP) to predict if a property would be removed from Airbnb's listings. I want to develop by using cutting-edge text mining methods to actual data consisting of property descriptions, host descriptions, and visitor comments.

Python and well-known libraries such NLTK, Scikit Learn, Keras, or PyTorch must be utilized in the project. I will build a powerful classification model with the use of multiple methods from NLP and the examination of various approaches.

1. DATA EXPLORATION

First, we had four datasets: "train", "train reviews", "test", and "test reviews." Both "reviews" and "test reviews" comprised several entries with the same index, I discovered after examining each dataset. I choose to aggregate the rows by index in both datasets to remedy this. After that, the combined data was stored as two new datasets, "train_reviews_new" and "test_reviews_new." These new datasets aggregated on the index but kept the original data and were unaffected by it.

The "train" dataset and the "train_reviews_new" dataset were merged using the index. Expanding the corpus for my models was my main objective. Making sure that the train and test datasets had the identical column names, I used the same methodology on the test dataset.

I made visualizations to look at the amount of data and the frequency of the most frequently used words in order to acquire insights into the data also visualization of amount od corpus in training set.

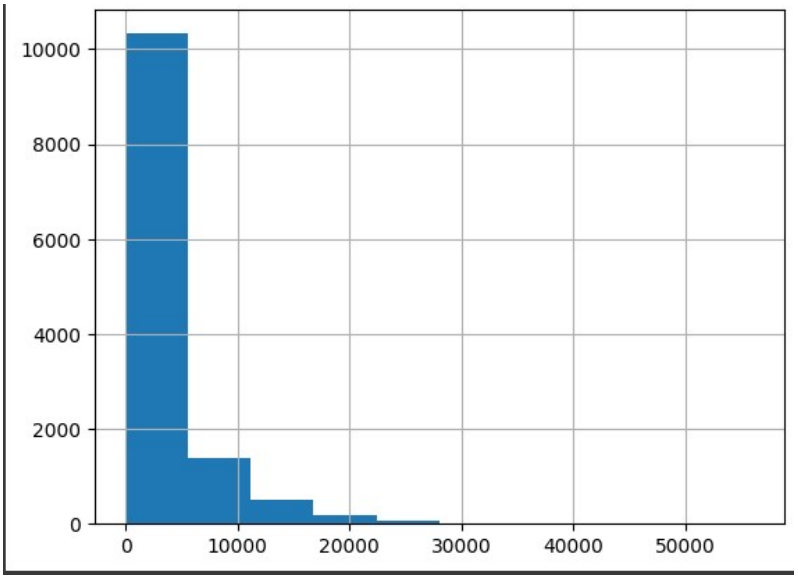

Fig.1 Words count in train corpus

2. DATA PREPROCESSING

Regarding preprocessing steps, I applied various techniques including the removal of stop words,regular expressions, puctoation, removel of emojies,lowercasing, lemmatization and stemming.Stopwords are any words that do not significantly add to the general meaning of an expression in any language. They can be freely ignored without affecting how the text should be understood.[1]Regular expressions are crucial for tokenization in text normalization. They normalize text, which helps analysis across a range of NLP tasks.[2]Removing punctuation. Emojis can be difficult for machines to interpret and may add unnecessary noise to your NLP modeling, and I decided to remove them.[3]Also I removed punctuation and numbers.Last part of preprocessing is lemmatization and stemming.One of the advantage of using stemming is improving model performormance.[4] A big advantage of lemmatization is accuracy because does not just breakdown of words, as stemming algorithms do[5].

3. FEATURE ENGINEERING

The first step I took was splitting my training corpus into a training set and an evaluation set. After that, I applied GloVe and TF-IDF embeddings. I did these two feature engineering techniques independently, and then I integrated them.The TF-IDF it can be defined as the calculation of how relevant a word in a series or corpus is to a text[6]. *GloVe Embeddings are a type of word embedding that encode the co-occurrence probability ratio between two words as vector differences.*[7]

4. CLASSIFICATION MODELS

First model which I applied was KNN,which stand for KNeighborsClassifier.It is one of the most commonly used technique and the idea of implementing is base on learning based on *the nearest neighbors of each query point k, where k is an integer value specified by the user.*[8]

Second model which I used was Logistic Regression,which is based on estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.[9]

---

[1] https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47 access at 12/06/2023
[2] https://web.stanford.edu/~jurafsky/slp3/2.pdf access12/06/2023 chapter 2
[3] https://towardsdatascience.com/primer-to-cleaning-text-data-7e856d6e5791 access at 12/06/2023
[4] https://www.datacamp.com/tutorial/stemming-lemmatization-python access at 12/06/2023
[5] https://www.datacamp.com/tutorial/stemming-lemmatization-python access at 12/06/2023
[6] https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/ access at 12/06/2023
[7] https://paperswithcode.com/method/glove
[8] https://scikit-learn.org/stable/modules/neighbors.html#classification access at 15/06/2023
[9] https://www.ibm.com/topics/logistic-regression access at 15/06/2023

The last model which I applied was MLP,which stand for Multilayer Perceptron.It is a type of neural network where mapping between inputs and output is non-linear[10].One of a biggest advantage of using MLP is flexibility, because can adapt to different text mining tasks.

5. EVALUATION AND RESULTS

According to the evaluation's findings, Logistic Regression (0.7032) had the greatest accuracy, followed by KNN (0.6584) and MLP (0.6244). However, when the F1-score, which balances precision and recall, was taken into account, MLP (0.4771), KNN (0.4838), and Logistic Regression (0.4341) obtained the greatest values.

Despite having the best accuracy, Logistic Regression may have trouble striking the correct balance between precision and recall, according to its F1-score. KNN outperformed Logistic Regression in terms of F1-score but behind MLP still. The performance of MLP was more evenly balanced in terms of precision and recall, yielding the greatest F1-score among the models that were tested and for that reason, predictions was made base on MLP model.

6. REFERENCES

- Daniel Jurafsky , James H. Martin.(2023), *Speech and Language Processing,Standford,chapter 2,p2*.[Accessed June 12 2023]
- SAI, Teja. Stop Words in NLP[onlie].June 10 2020.[Accessed June 12 2023] Avalibe from: https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47
- SEUNGJUN, Kim. Primer on Cleaning Text Data. Sep 2, 2022 [Accessed June 12 2023] Avalibe from: https://towardsdatascience.com/primer-to-cleaning-text-data-7e856d6e5791
- KURTIS ,Pykes. Stemming and Lemmatization in Python. Feb 2023 [Accessed June 12 2023] Avalibe from: https://www.datacamp.com/tutorial/stemming-lemmatization-python
- Riturajsaha. Understanding TF-IDF (Term Frequency-Inverse Document Frequency). Feb 2020 [Accessed June 12 2023] Avalibe from: https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. [Accessed June 12 2023] Avalibe from: https://paperswithcode.com/method/glove

[10] https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141 access at 15/06/2023

- Scikit-learn Nearest Neighbors Classification [Accessed June 15 2023] Avalibe from: https://scikit-learn.org/stable/modules/neighbors.html#classification
- What is logistic regression? Accessed June 15 2023] Avalibe from: https://www.ibm.com/topics/logistic-regression
- Carolina, Bento. Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis. Sep 21, 2021, [Accessed June 15 2023] Avalibe from: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141

7. APPENDIX

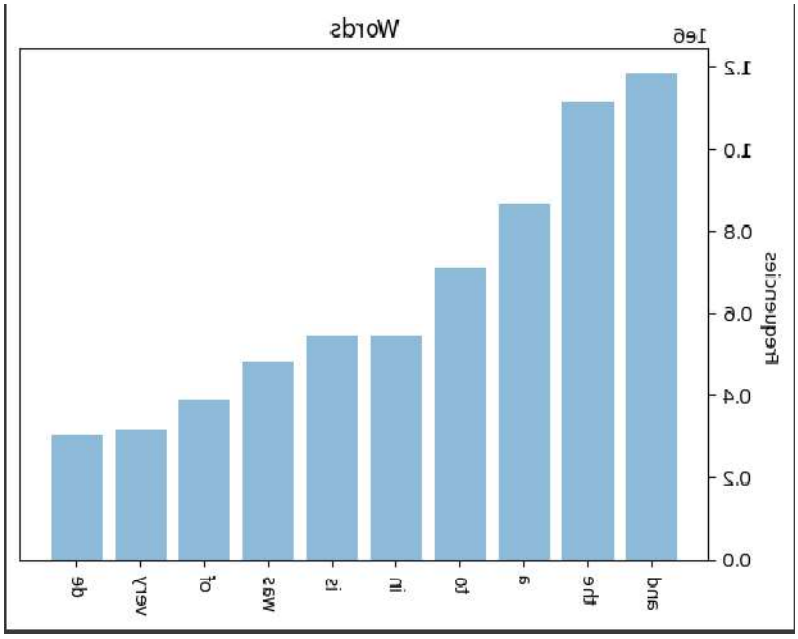

Fig.2 Words frequency