Sentiment Analysis of Reviews and Business-User Clustering on YelpDataset

Artificial Intelligence Project

Paola Persico, Igor Iurevici, Salvatore Fiorilla

October 5, 2020

Università degli Studi di Bologna

Overview

Introduction

Introduction

In this project we've worked with the **Yelp's academic datasets** available on their website. In particular we consider the following datasets:

- **Review** (5.89 GB)
- **Business** (145 MB)
- **User** (3.04 GB)

Datasets documentation source:
https://www.yelp.com/dataset/documentation/main

Goal

We have three different notebooks, each with a specific **goal**:

- Yelp review sentiment analysis
- Description-based business clustering
- Preferences-based user clustering that exploits the previously determined business' clusters

We developed in a **Colab environment** in order to exploit greater computational capabilities than a regular computer.
Nonetheless we had to work with smaller chunks of data due to RAM limits.

Models

Models

- Review Sentiment Analysis and performance evaluation
- Description-based Business Clustering model
- Preferences-based User Clustering model

Review sentiment analysis

Several models were built and evaluated in order to find the best one to **automatically classify reviews in positive and negative** ones based on the text' features.

Notebook link:
https://colab.research.google.com/drive/1BtYiHLLQ7iD5tB1KqXH5HnsXro9Le3Uj?usp=sharing

Approach

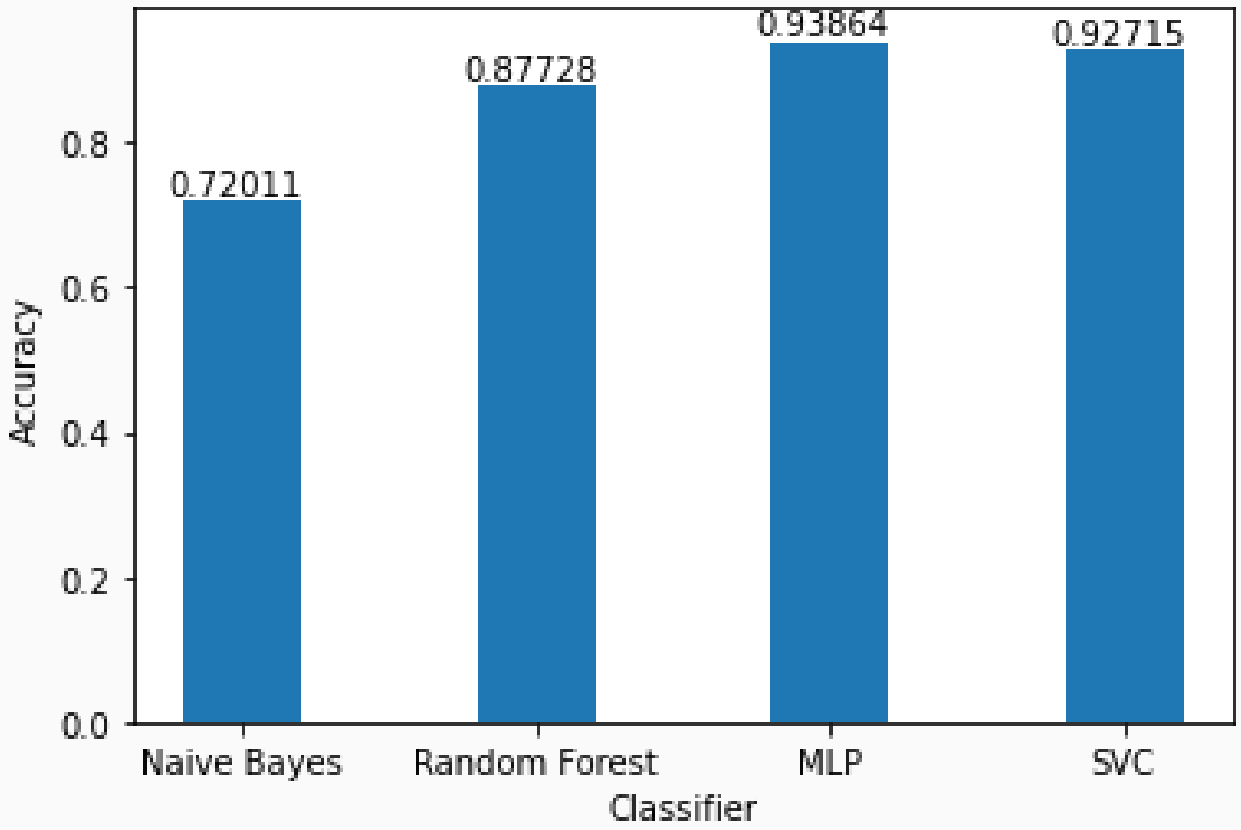
word2vec was used during the word-embedding phase.

Most famous classification models were used and evaluated:

- Naive Bayes
- Random Forest
- Multi-Layer Perceptron
- Support Vector Classifier

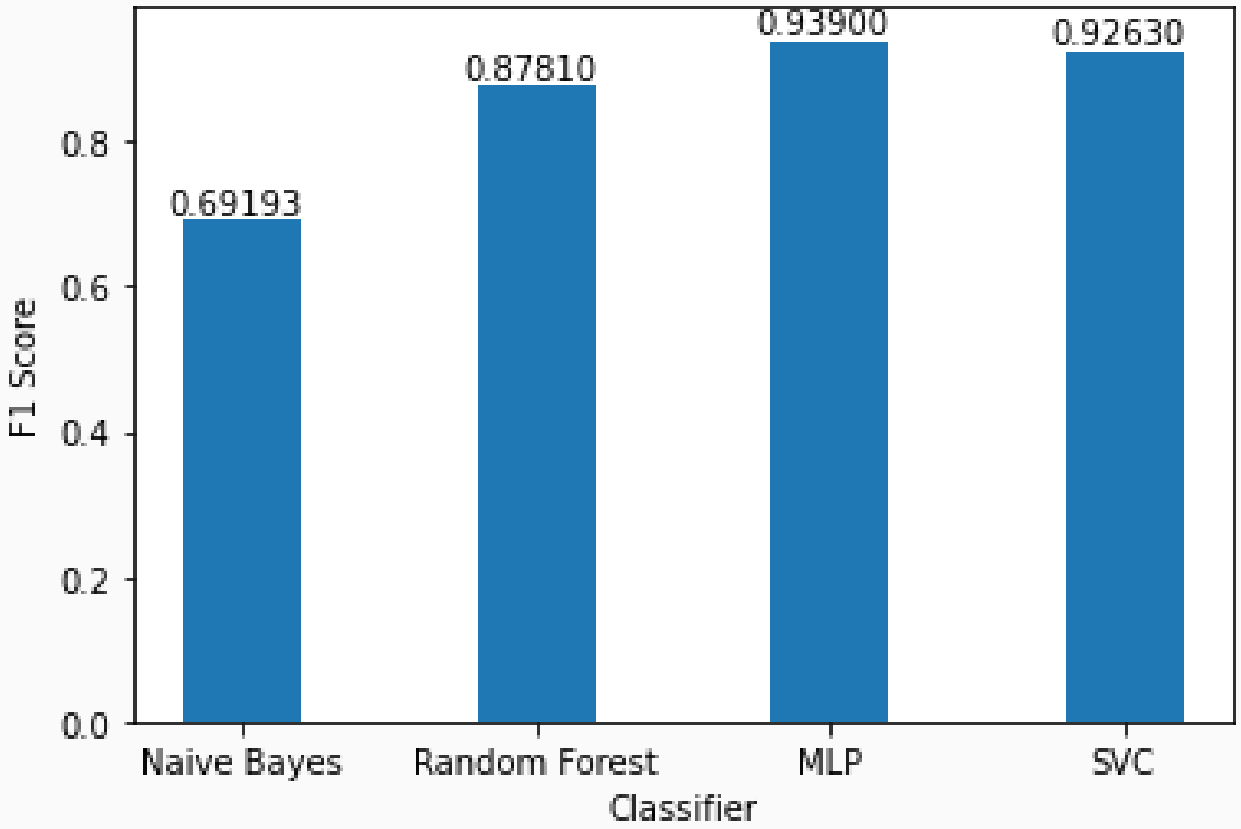
Evaluation 1/3

- Accuracy:



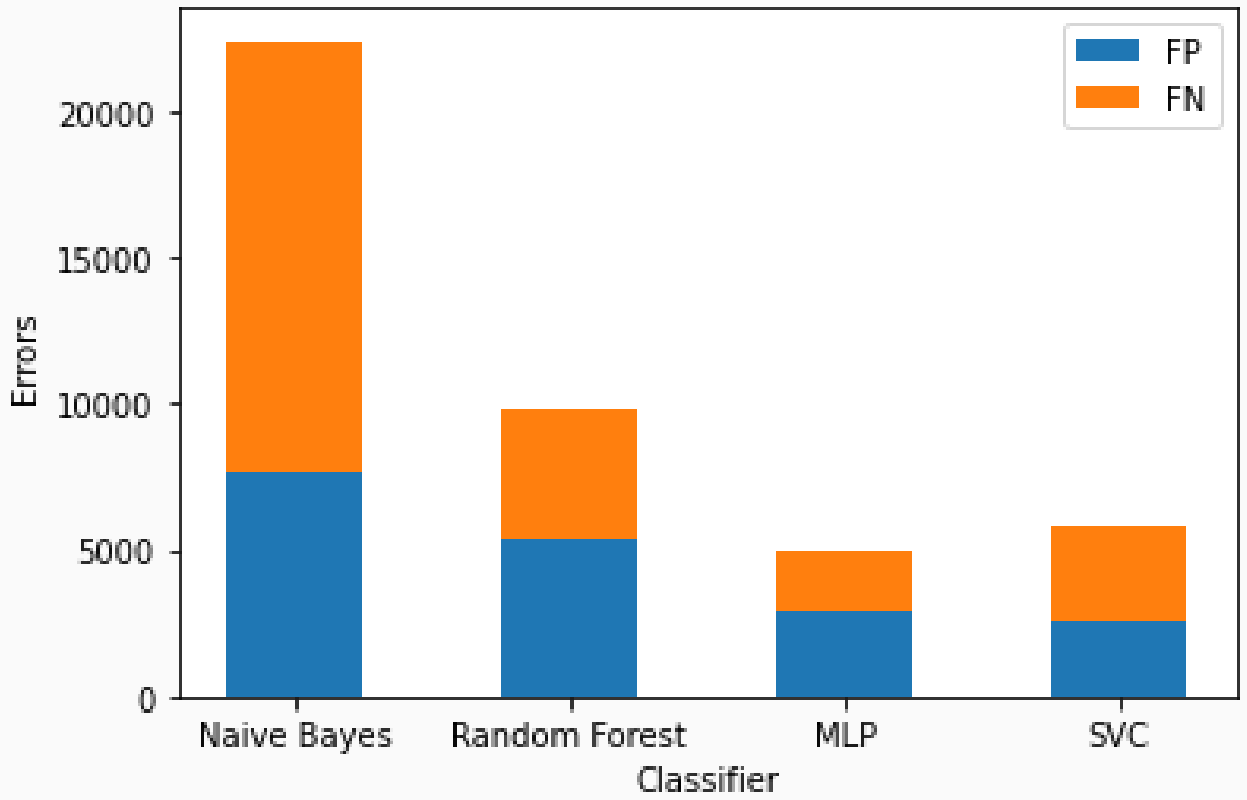
Evaluation 2/3

- F1 score:



Evaluation 3/3

- Errors:



Models

- Review Sentiment Analysis and performance evaluation
- Description-based Business Clustering model
- Preferences-based User Clustering model

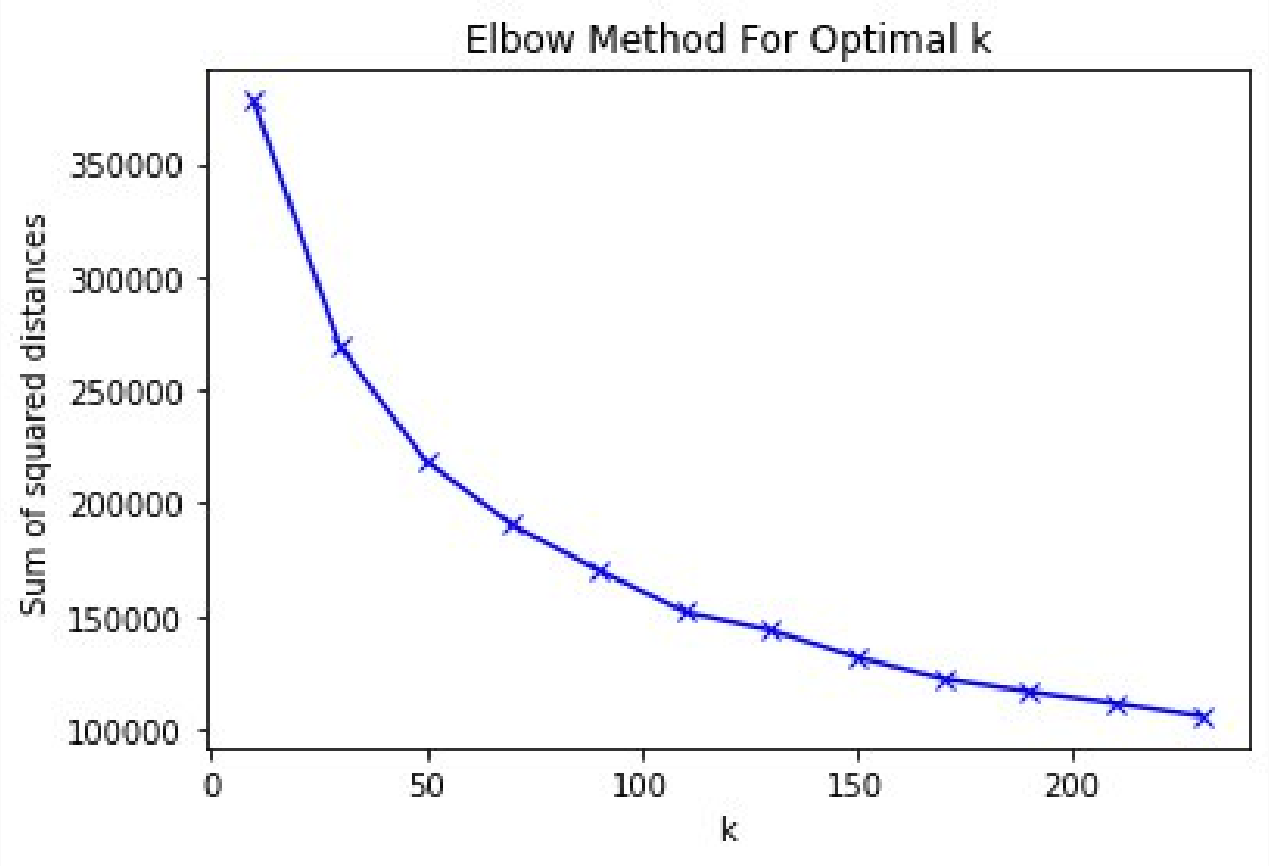
Description-based business clustering

In order to automatically group businesses according to their categories, **K-means clustering algorithm** was used.
A **city-based filter** is implemented and allows to filter clusters according to the chosen location.

Notebook link:
https://colab.research.google.com/drive/1rRFCmXHqfu337tBd7zBPmymyLStxDzJw?usp=sharing

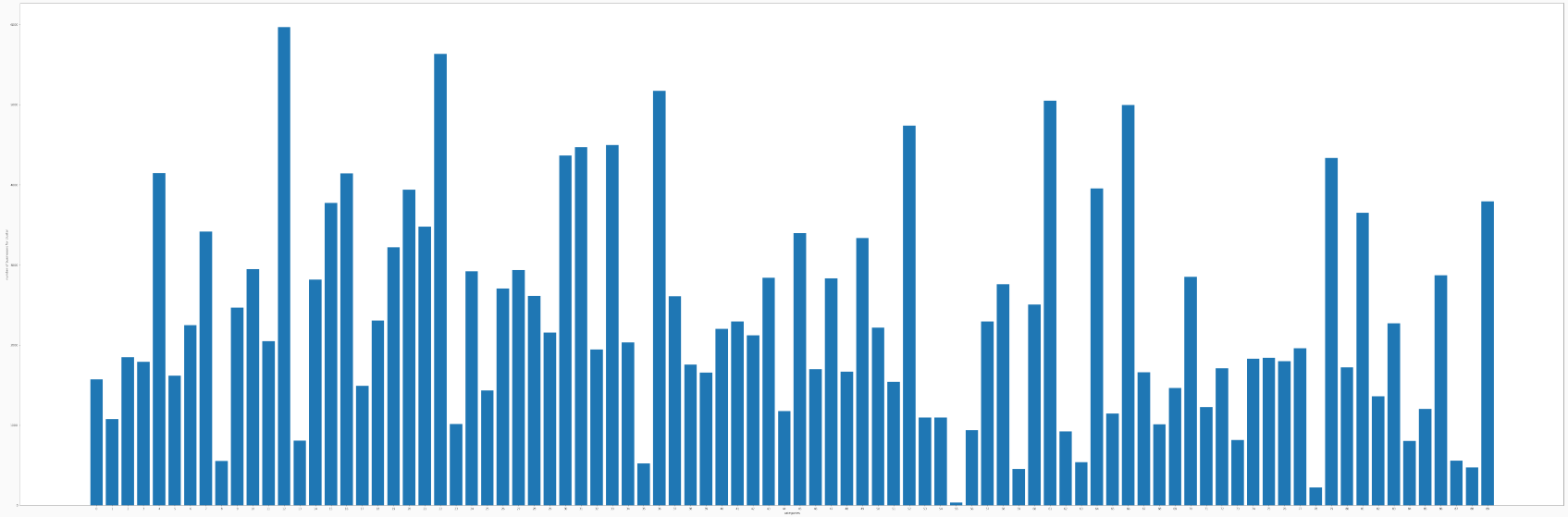
Elbow

- Elbow:


Elbow Method For Optimal k
Sum of squared distances
350000
300000
250000
200000
150000
100000
0
50
100
150
200
k

Distribution

- Businesses categories distribution:



x-axis: Clusters

y-axis: Businesses n.

Models

- Review Sentiment Analysis and performance evaluation
- Description-based Business Clustering model
- Preferences-based User Clustering model

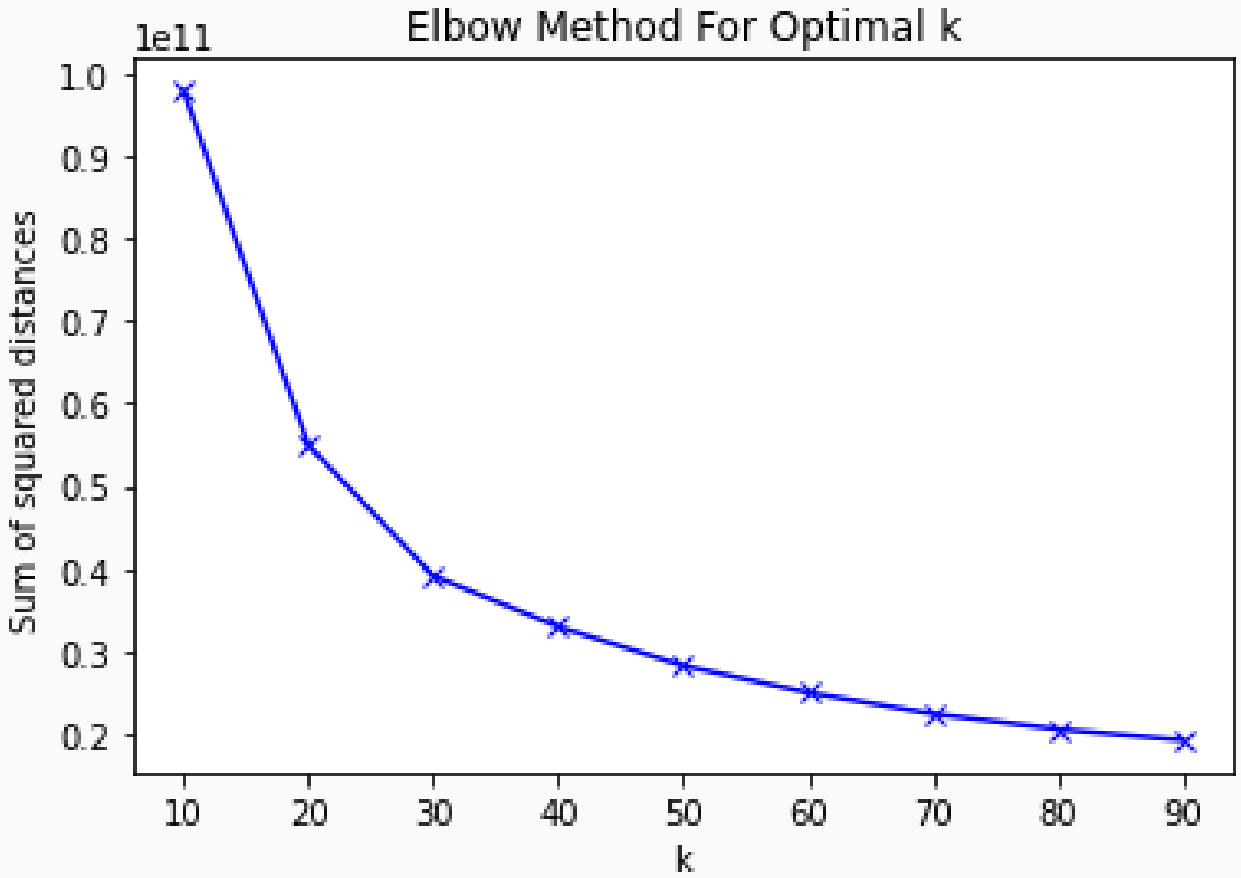
Preferences-based User Clustering

A further grouping was made on the user dataset.
Heavier pre-process operations were executed in order to define a new user dataset that includes for each user the **reviews' number** and **average stars** related to every business' cluster previously defined.
Afterwards a K-means algorithm was again performed to define users similarities according to their involvement in every business' group.

Notebook link:
https://colab.research.google.com/drive/1hNqLe7gWIb-3kH5axxNUGRlAeHjdPdQr?usp=sharing

Elbow

- Elbow:


Elbow Method For Optimal k
1e11
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Sum of squared distances
10
20
30
40
50
60
70
80
90
k

Conclusion

Conclusion

- The **accuracy and precision metrics** of the review sentiment analysis model are definetly satisfying;

- Grouping businesses according to their similar categories allows us to visualize their distribution in a specific location and may be fully **exploited in a spatial representation**;

- Joining multiple relevant datasets gave us an overall clear idea about the distribution of the users groups according to their **involvement** in different business categories.

Thanks for the attention!