# An NLP Analysis on the Reviews from De Kas Restaurant Amsterdam

Data Analysis Bootcamp May 2021
IronHack

Laura Trapero

# Contents



01

# Project description

- Understand the customers’ feedback and know what are the strengths and weaknesses from the business from their eyes.
- Use NLP to extract quick and valuable insights from the customers’ reviews

## Goals

### 1

**Which features are more valuable for the costumers.**

### 2

**Which areas improve to deliver a better experience for the costumers.**

### 3

**Establish a model to determine whether a review is positive or negative, and which features are important on the decision.**

02

# Context

- Sustainable restaurant
- 2001
- Michelin star
- Located in a park, in a greenhouse.
- They have gardens in Amsterdam and the Beemster.
- Fresh own products drown in open air, greenhouse and hydroponically.
- Local farmers
- Seasonal food

03

# Workflow

## Getting the data

- Trip advisor Web Scraping - Selenium
- 1038 reviews
- Features:
  - Id
  - Date
  - Header
  - Score
  - Review

## Cleaning data

Clean text for NLP methods:

- remove special characters
- remove extra spaces
- remove digits
- remove urls's
- convert to lower case

## Pre-processing

- Create a corpus
- Tokenization
- Lemmatization
- Remove non-meaningful words

- Classify reviews in:
  - positive reviews > 3
  - negative reviews < 3

- Down & up sampling

## Words Analysis

- Words frecuency
- TF-IDF ( Term Frequency-Inverse Dense Frequency) / ngrams (2,3)

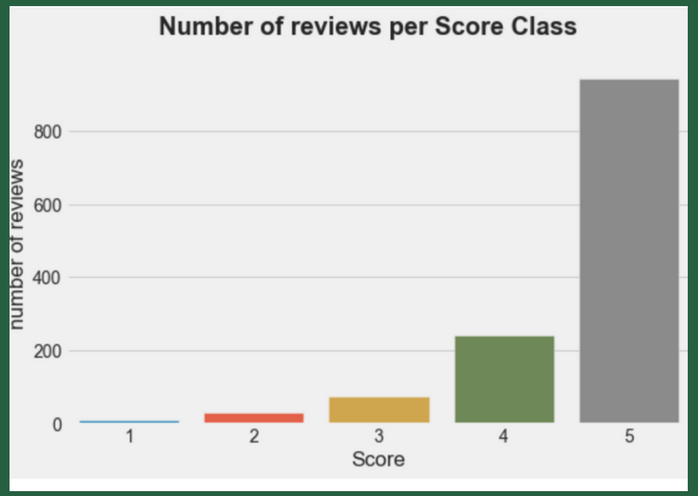
## Models

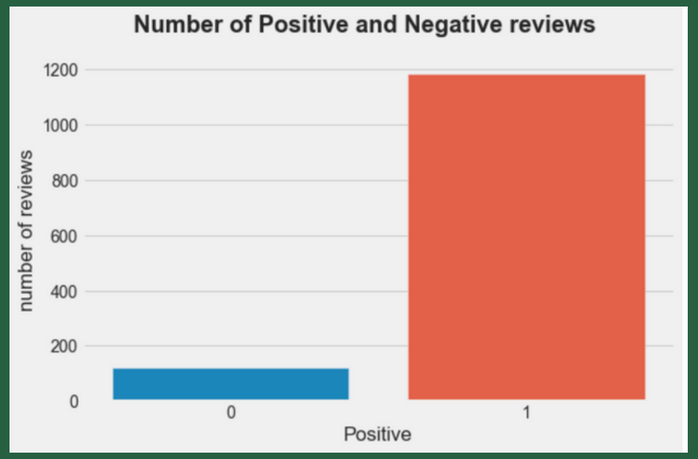
- Classification models

04

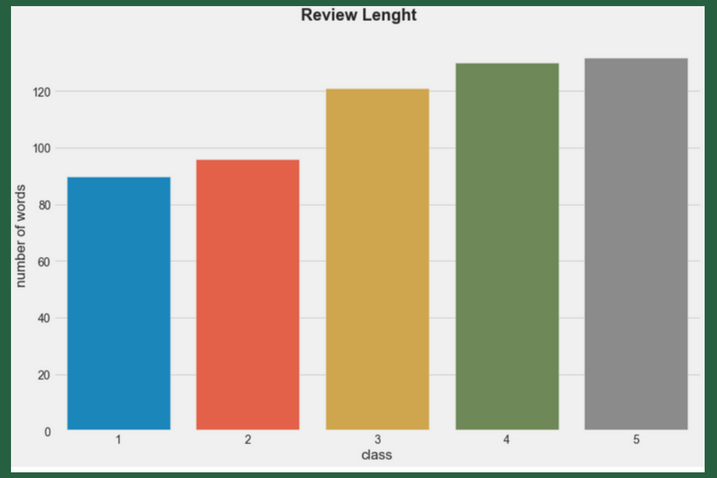
# Exploratory Analysis

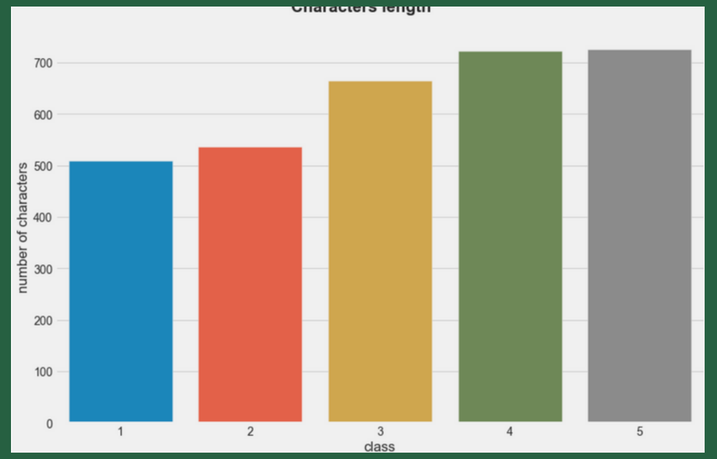
## How many reviews per score class?




## Calculated metrics:



04

# Bag of words


Positive

Negative

## TF-IDF ( Term Frequency- Inverse Dense Frequency) / ngrams )

By evaluating TF-IDF:  number of “the words used in a sentence vs words used in overall document”, common theme in all the documents (*sort by tif-idf score/weight)/20 features/2g-3g

### Negative

**appetizer tasteless**
fresh ingredient
three course
left hungry
**guinea fowl**
bread tasty
three starter
dining experience
**fish** bland
**course soup**
bad service
course fish
**high expectation**
paid euro
bottle wine
first course
set menu
main course

**pork** chorizo samphire
pork veal preferred
**portion extremely tiny**
poor experience want
**poor selection cheese**
portion extremely small
pork main accompanied
pork medallion special
pork rib small
**salad** main course
choice first course
**euro per person**
main course tiny
looking forward dinner
first course soup
waiter wearing mask

### Neutral

**brussel sprout**
paid euro
vegetable dish
small plate
**small piece**
six course
service good
nothing special
lovely setting
ice cream
fresh vegetable
small portion
wine pairing
**set menu**
green house
tasting menu
course menu
first course
main course

**potato** crisp brought
possibly best thing
take long time
potential management act
**potatoe brocolli flowe**r
potato thought unusual
**potato nothing extraodinary**
good great wait
menu per person
main course **chicken**
**lamb** main course
**hour main course**
review guide book
serf **daily** menu
six course tasting
greenhouse middle park

### Positive

**service good**
best meal
excellent service
**vegetable grown**
grown site
next time
highly recommended
course meal
**service excellent**
**staff friendly**
**tasting menu**
fixed menu
**wine list**
well worth
dining experience
**green house**
main course
highly recommend
**wine pairing**
**set menu**

greenhouse middle park
**easy tram ride**
**vegetable grown site**
set menu course
good wine list
euro per person
**wine pairing menu**
book well advance
best meal ever
best dining experience
home grown vegetable
service attentive friendly
made feel welcome
staff friendly helpful
course wine pairing
back next time
well worth trip

05

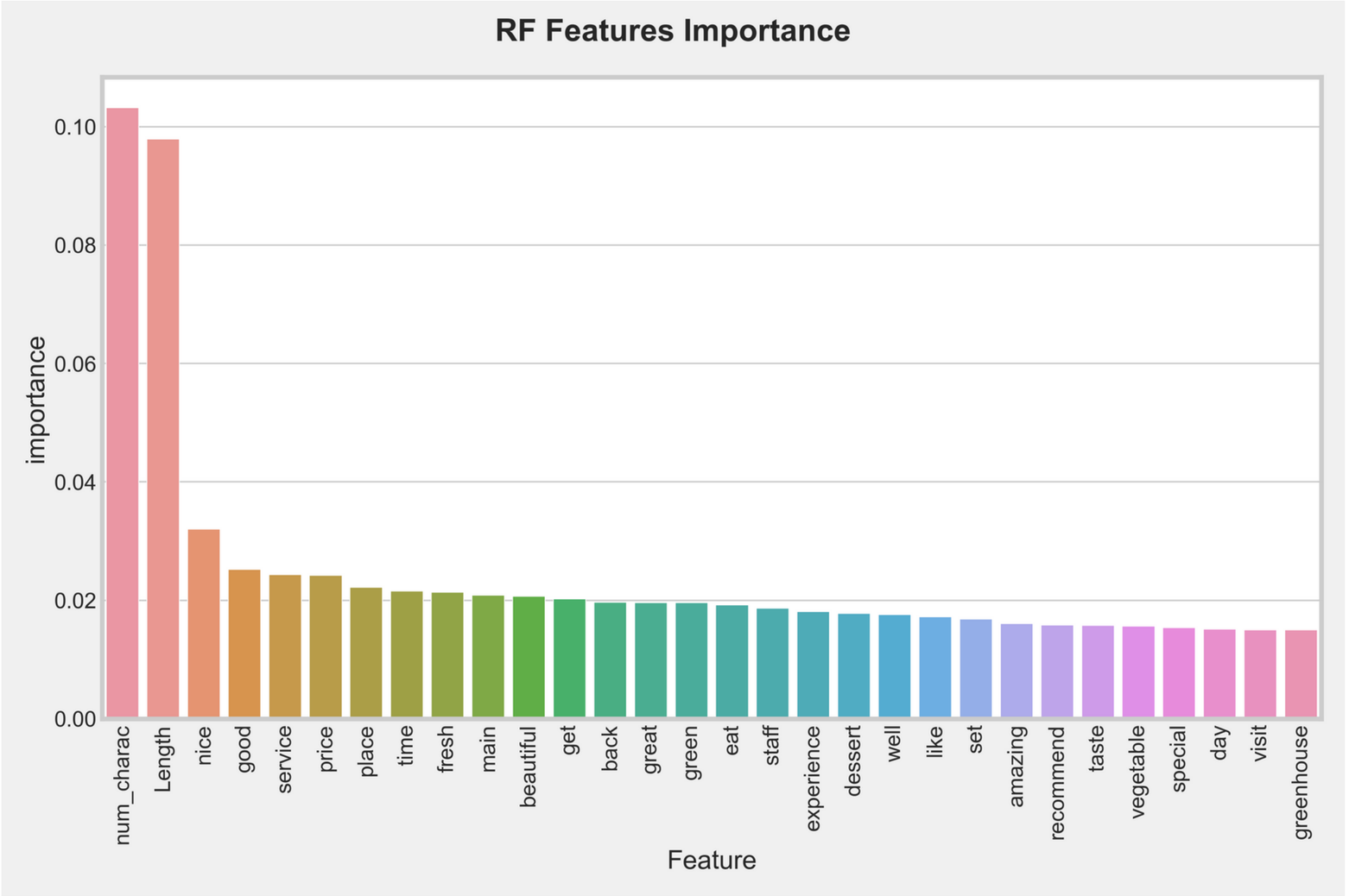
# Modeling

But training a model to predict this rating will help us find which words (features) are key for customers.

- Down & up sampling (train set)
- Dummify set of words
- Models (balanced, unbalanced, 5 categories, 2 categories, 1-2-3 g,

  - Random Forest (extraction features)
  - Logistic Regression (RobustScaler)

- Bad predictions, with really low kappa values
- Predicts better the majority class



06

# Conclusions and recommendations

- Highly valued by clients: service, location, building, own garden, wine pairing, set menu

- Rethink the use of meat and dairy, portioning, appetizers, starters (cheeses, soup, salads) and the way to use basic ingredients like potatoes and cabbage family vegetables.

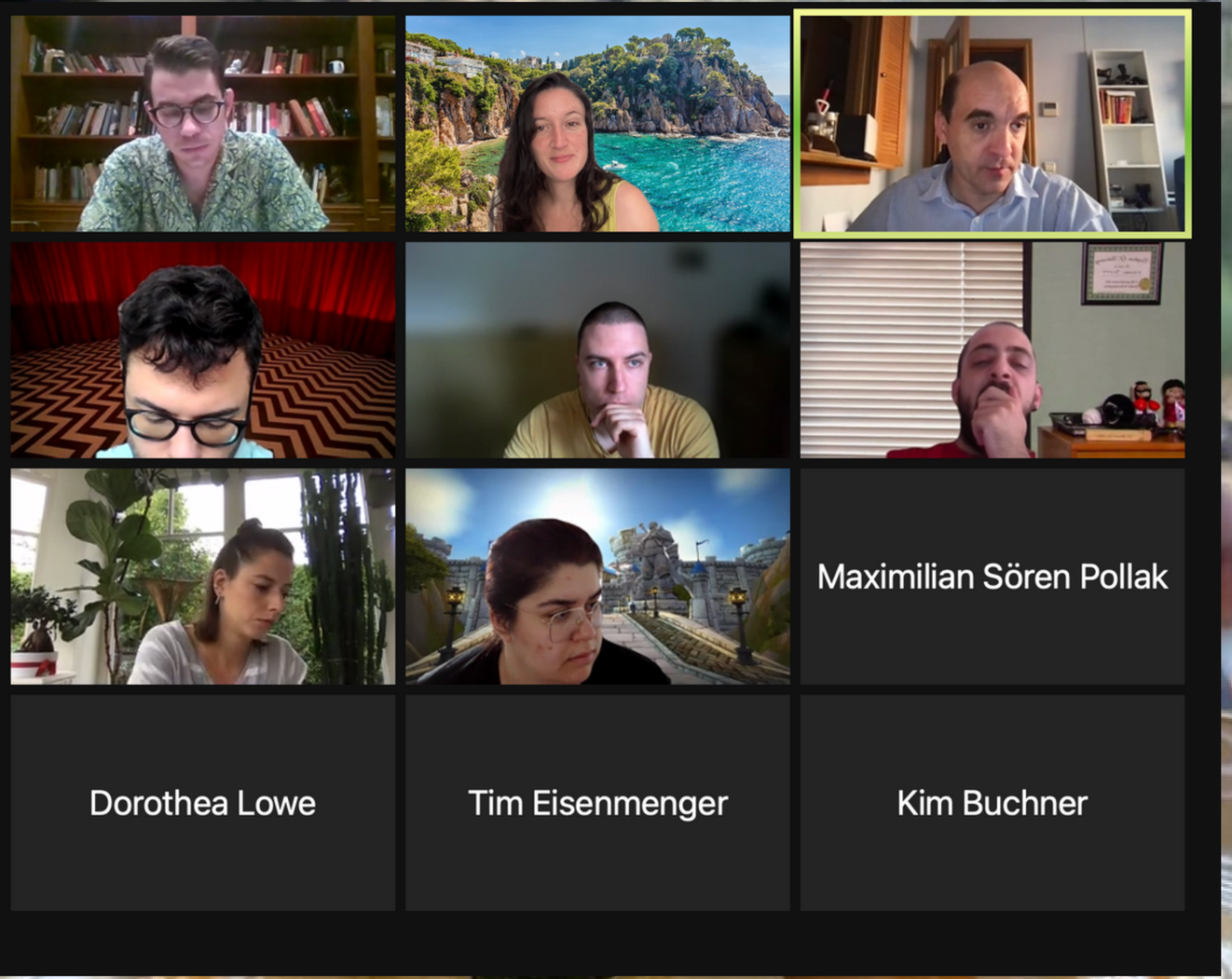
- Not good predictions of the models, predicts better the majority class.

07

# Future Improvements

- Get more reviews from other sites like Yelp, google and local ones.
- Use more complex models and NN
- Time Series Analysis
- Themes Impact Analalysis
- NLP on the Reviews Title
- Embeding methods: represent reviews as vectors representation to be able to apply clustering algorithms to detect topic and other methods like Word2Vec Architecture Implementation (word representations in Vector Space )
-

# Thank you!




Questions?

or it is too early?