Università degli Studi di Milano – Bicocca
Master's Degree in Data Science
Academic Year 2022/2023

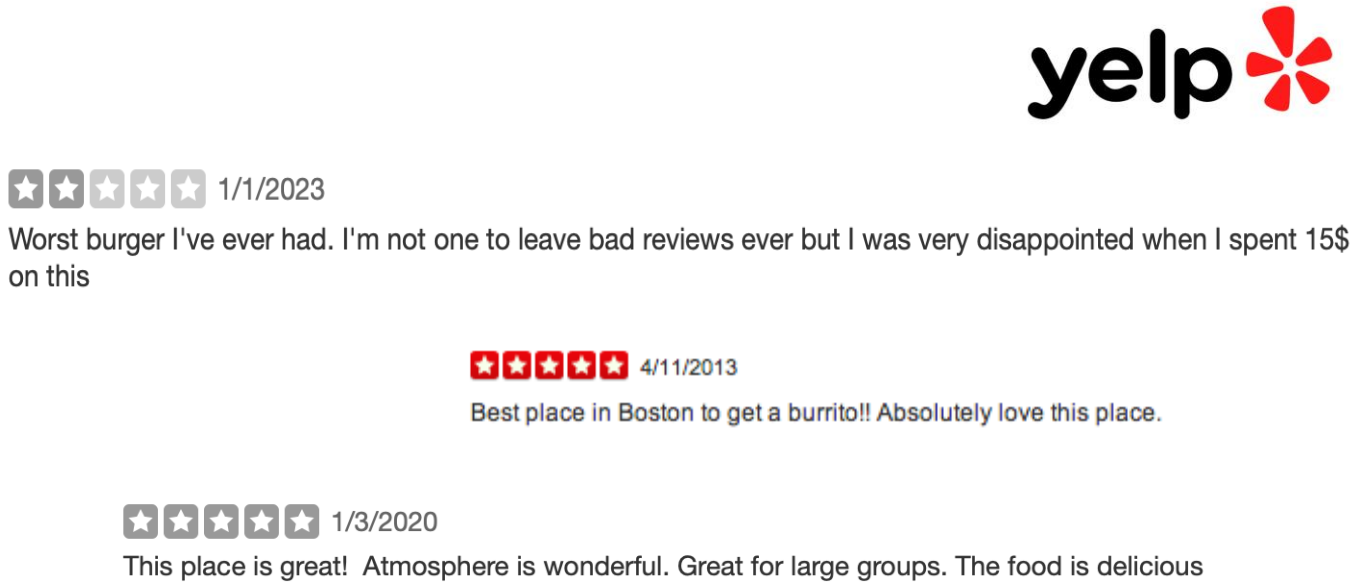
Text Mining & Search Project

Analysis of Yelp reviews

Julia Lan Bui Xuan, 882385
Alessio Pasinato, 887000

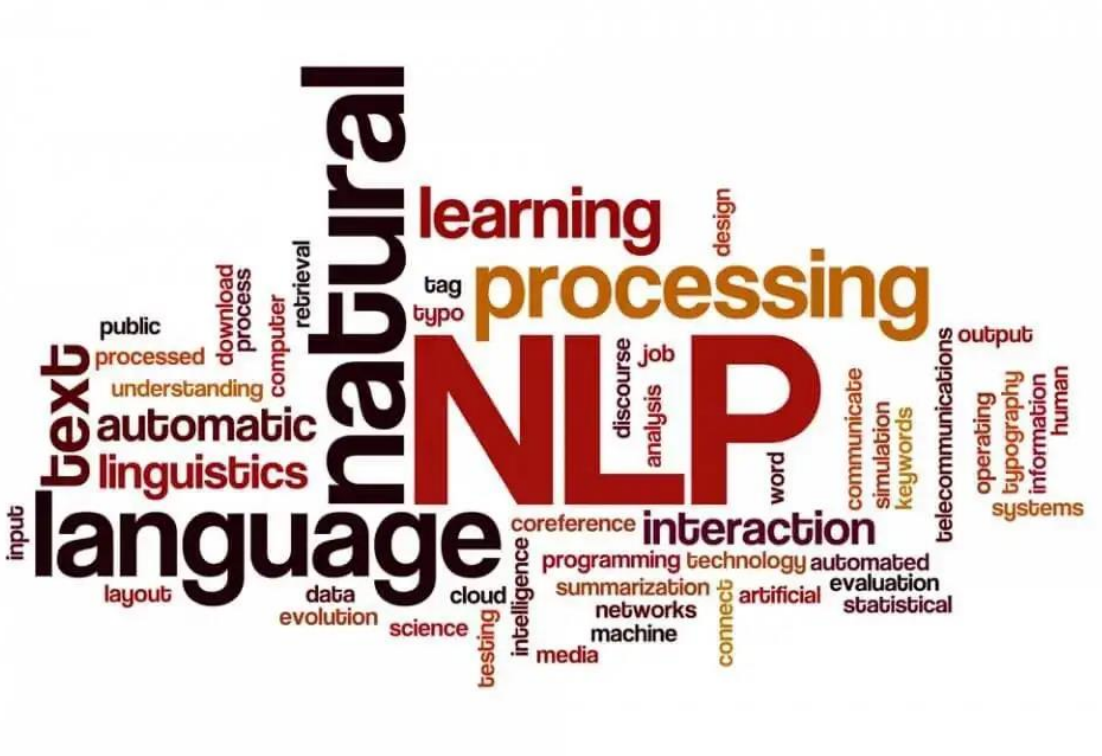
Introduction to Yelp

- Online platform, founded in 2004 and headquarted in California
- Used by individuals to **rate** and **review** local businesses
- Provides a feature for businesses to claim their listing and respond to reviews
- A dataset consisting of **6,990,280 reviews** and **150,346 businesses**



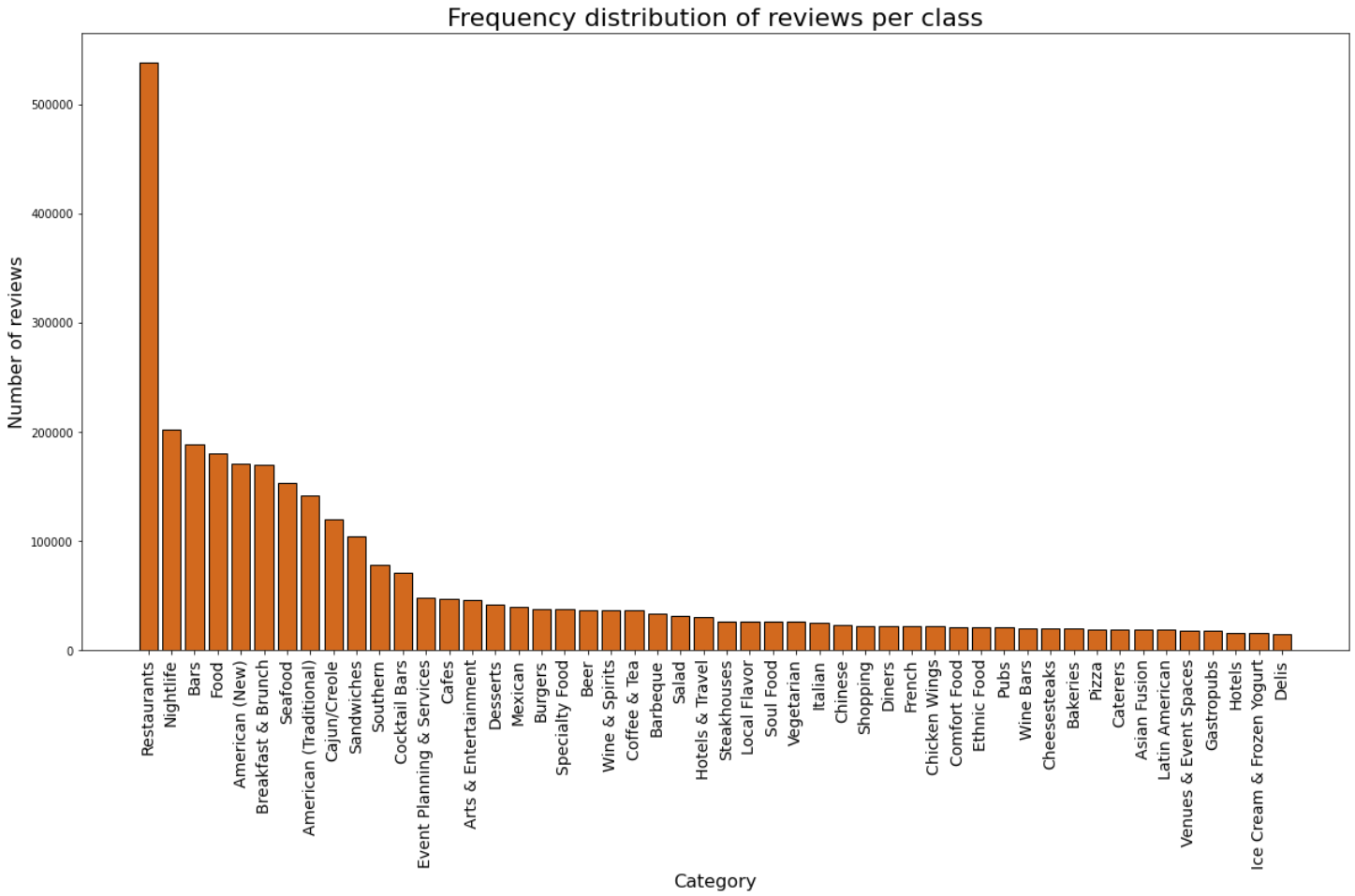
Objectives

- Development of a model capable of **labelling reviews**
- Comparison of **different text representations**' performances over the classification task
- Performance **topic modelling** to detect the most discussed topics in the reviews



Data exploration

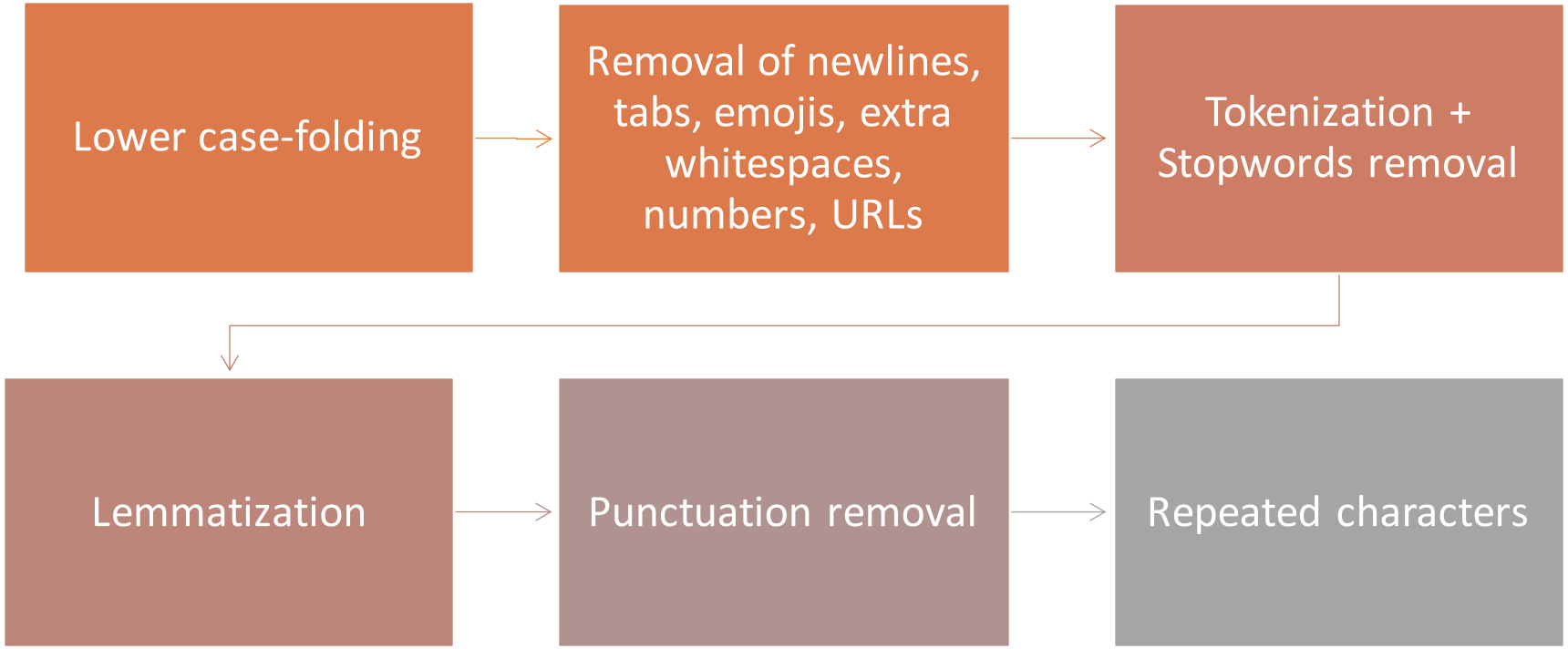
- N. reviews: 570,438
- N. classes: 200
- Mean reviews per class: 19,150
- Least represented: "Hot Dogs", 1,045
- Most represented: "Restaurants", 538,450



*Histogram for **top 50 classes***

Text Pre-processing

Necessary **steps for cleaning and preparation of raw text data** for further analysis



Text classification

Multilabel multiclass classification

- 15 randomly selected businesses
- “Bars”: 14383 occurencies
- “Donut”: 1074 occurrencies
- **Stratified sampling technique**: 70% for training, 30% for testing
- OneVsRest classifier

27,512 reviews
44 classes

TF-IDF representation

TF-IDF:

- Removed words present in less 5 documents
- 500 features
- WordNet lemmatizer
- SpaCy lemmatizer

Doc2Vec:

- Vectors' size = 50
- 100 epochs

	Precision	Recall	f1-score
micro avg	0.93	0.75	0.83
macro avg	0.94	0.73	0.82

Table 1. Scores WordNet lemmatizer

	Precision	Recall	f1-score
micro avg	0.93	0.75	0.83
macro avg	0.95	0.75	0.83

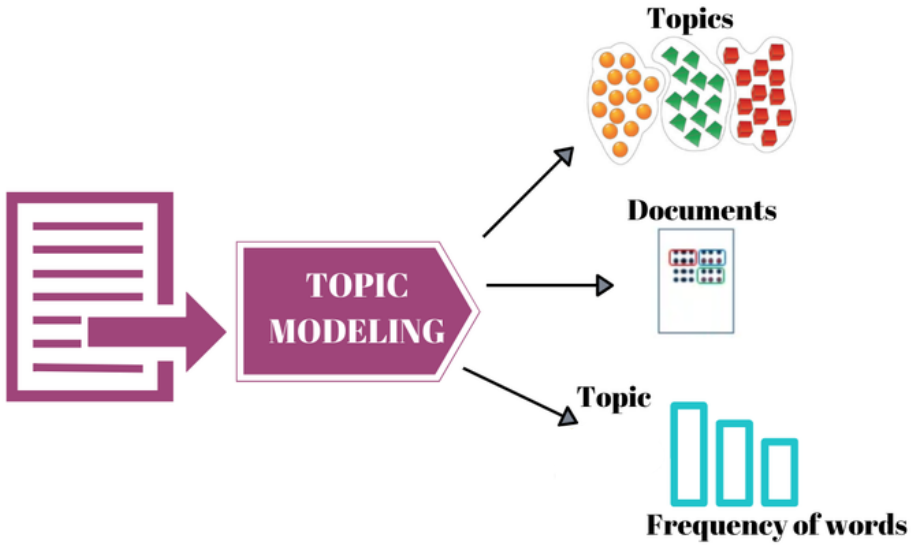
Table 2. Scores SpaCy lemmatizer

	Precision	Recall	f1-score
micro avg	0.83	0.55	0.66
macro avg	0.81	0.53	0.62

Table 3. Scores Doc2Vec

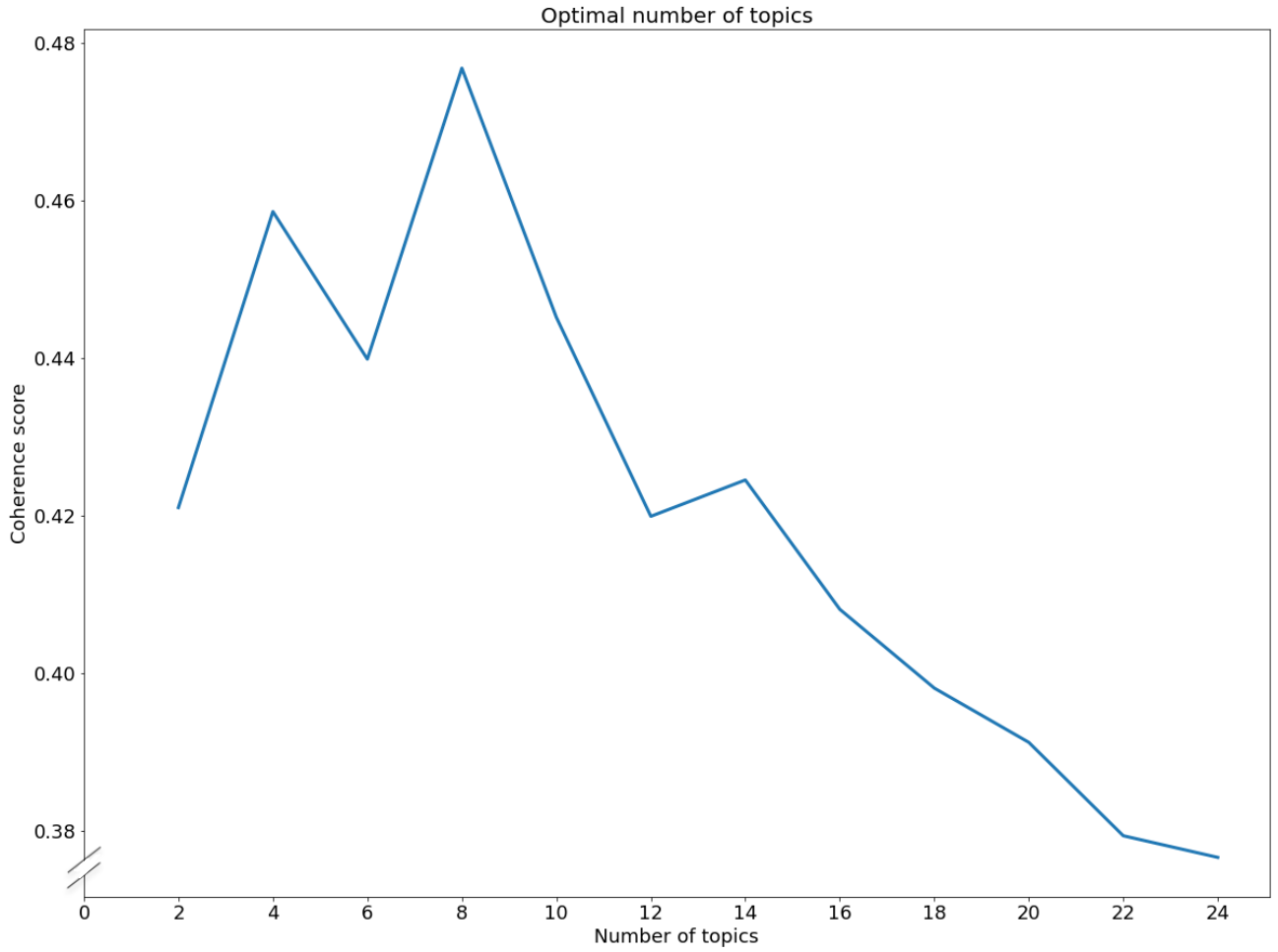
Topic Modeling

- **Latent Dirichlet Analysis** (LDA)
- **Bag of Words** assumption
- **Term Frequency** (TF) representation
- **Removal** of too frequent and too rare words
- Evaluation in terms of:
  - **Perplexity:** measuring uncertainty
  - **Coherence**: measuring the degree of semantic similarity between high scoring words in the topic
  - **Human judgment**: observing the top ‘n‘ words in a topic, word intrusion



Optimal number of topics

- **K = 8**
- Coherence score = 0.477
- Using a lower number of topics, allowed to gain **more understandable interpretations** of the results, compared to a larger number of topics.



Optimal number of topics. Truncated Y axis. Coherence scores against number of topics

Topics extracted

1. Seafood
2. Hotel
3. Dessert
4. Breakfast / Brunch
5. Bar
6. Sandwiches
7. Positive food experiences
8. Orders / Waiting time

Wordclouds for each of the 8 topics extracted.

Conclusions

Classification

- Equivalence between WordNet and SpaCy lemmatizer
- Better performance of tf-idf over Doc2Vec
- Classificator:
  - 93% precision
  - 74% recall

Topic modelling

- LDA: 8 food-related extracted topics