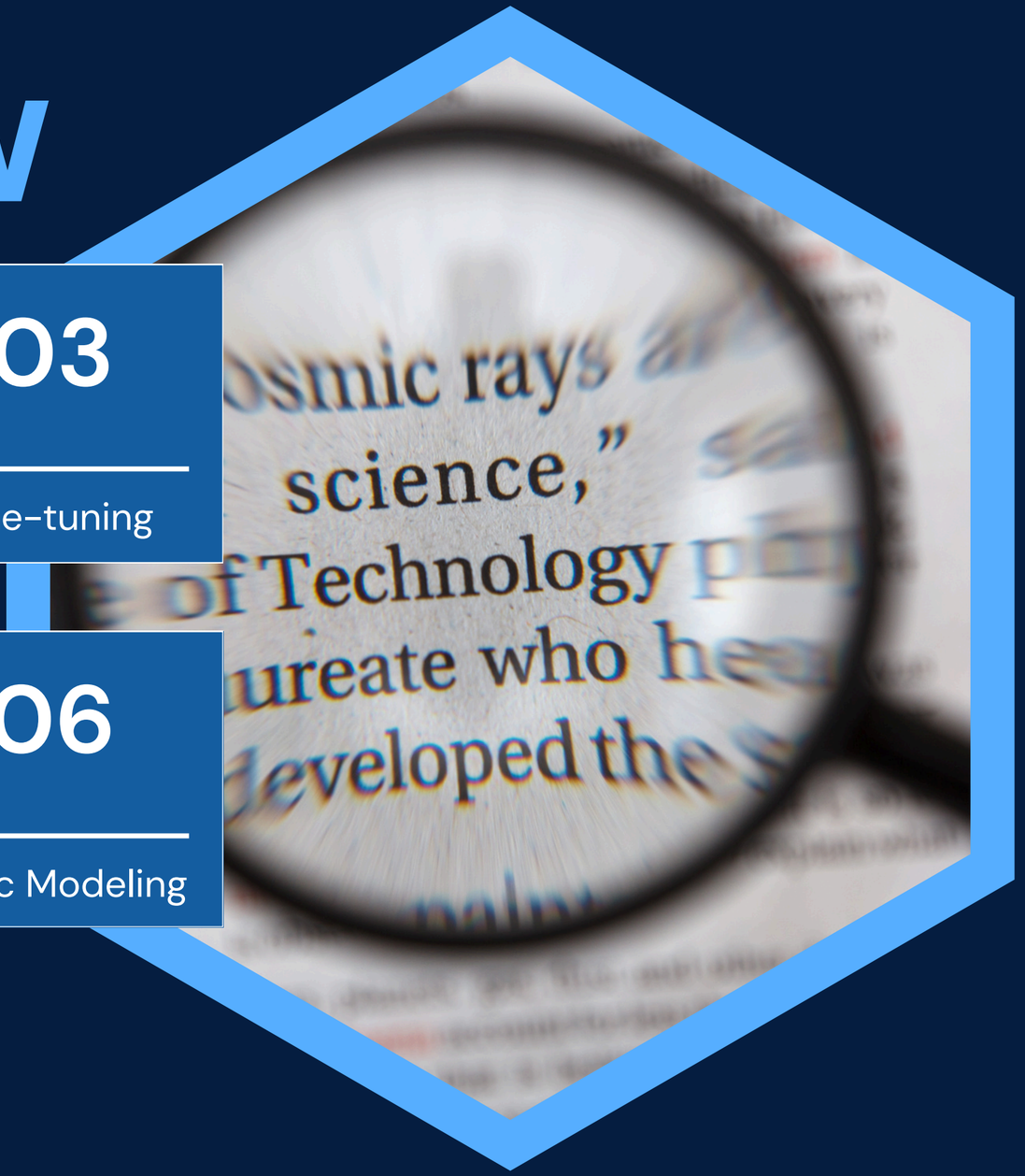
TEXT MINING PROJECT

Presented by:
Andrea D'Amicis 869008
Gabriele Sormani 869217
Alisia Sallemi 866453

OVERVIEW

01	02	03
Dataset	Preprocessing	Fine-tuning
04	05	06
Text represent.	Clustering	Topic Modeling



DATASET

Yelp dataset is real-world data related to businesses including reviews, photos, check-ins, and attributes like hours, parking availability, and ambience.
To carry out this project, the data related to reviews and business are considered and merged taking only relevant categories.

86522
Fine tuning

92376
Test

Category	Description/Notes
Bars	Bars and pubs for drinks and aperitifs
American (Traditional)	Traditional American cuisine
American (New)	Modern American cuisine
Breakfast & Brunch	Breakfast and brunch
Mexican	Mexican cuisine
Beauty & Spas	Beauty and wellness centers
Shopping	Shops and malls
Seafood	Seafood restaurants
Pizza	Pizzerias
Sandwiches	Sandwich shops
Italian	Italian cuisine
Automotive	Automotive services and parts
Coffee & Tea	Coffee shops and tea houses
Burgers	Burger restaurants
Hotels & Travel	Hotels and travel services
Home Services	Home services
Ice Cream & Frozen Yogurt	Ice cream and frozen yogurt shops
Sushi Bars	Sushi restaurants
Health & Medical	Health and medical services

PREPROCESSING

01 **Normalization**: Removes non-alphabetic characters and converts text to lowercase.

02 **Stopword Removal**: Filters out common words to retain meaningful terms.

03 **Tokenization**: breaks a stream of text into meaningful units, called tokens.

04 **Lemmatization**: Breaks text into words and reduces them to their base forms. This process was guided by Part-of-Speech (POS) tags

05 **Output**: Produces cleaned, standardized text ready for NLP or machine learning tasks.

FINE TUNING

To improve the ability of the Sentence Transformer model all-MiniLM-L6-v2 and capture the semantic relationship in the reviews, the fine-tuning of the model is performed. The settings of the training are:

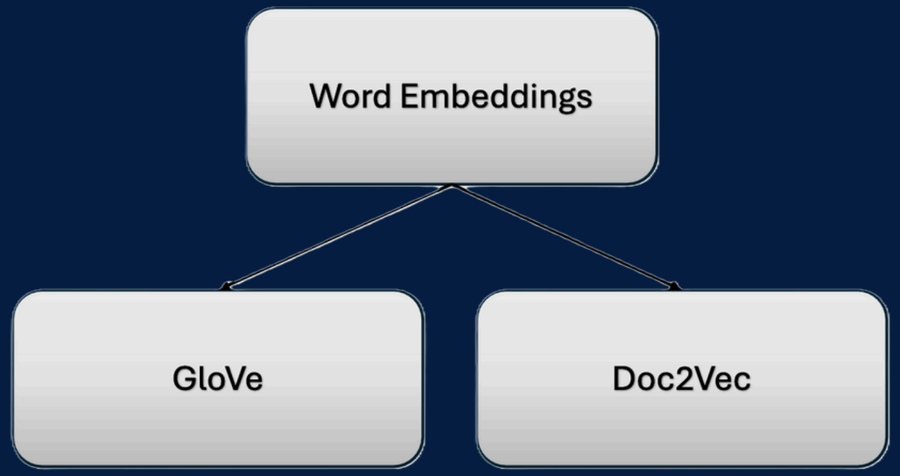
- Train instances: 40000
- Loss function: BatchTripletLoss
- Learning rate: 0.00002
- Epochs: 3

	Label	Text
Anchor	7	I'm one of their regulars and the food is pretty consistent. Their BBQ chicken pizza with thin crust and Chicken Alfredo pasta are amazing! We order from them weekly and delivery is pretty fast. I don't know why the reviews are so low because they never disappoint me.
Positive	7	Never have left a restaurant so upset and disappointed in a experience. Took over an hour for food on a Monday night with a half empty restaurant. No apology no excuses, just a simple lack of caring for the service of their guests. Food was good, but it's not worth the long wait, lack of service, and utter disregard for guests. I'd take my business elsewhere.
Negative	3	We stayed here for a long weekend last year. 2 kids, 1 a toddler and we had a great time. All the restaurants were good. My son and husband had fun at the arcade. Hubs enjoyed gambling. My favorite hotel in Reno.

TEXT REPRESENTATION

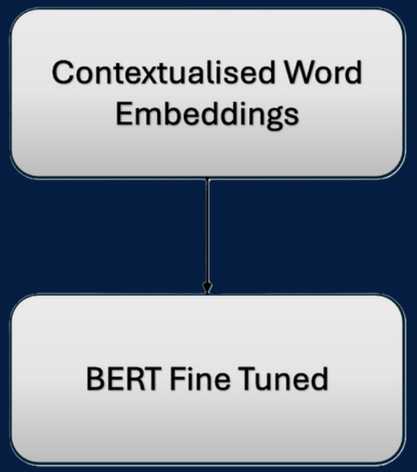
Text Preprocessing:

- Normalization
- Stopwords Removal
- Tokenization
- Lemmatization



Text Preprocessing:

- Normalization
- Stopwords Removal
- Tokenization



Text Preprocessing:

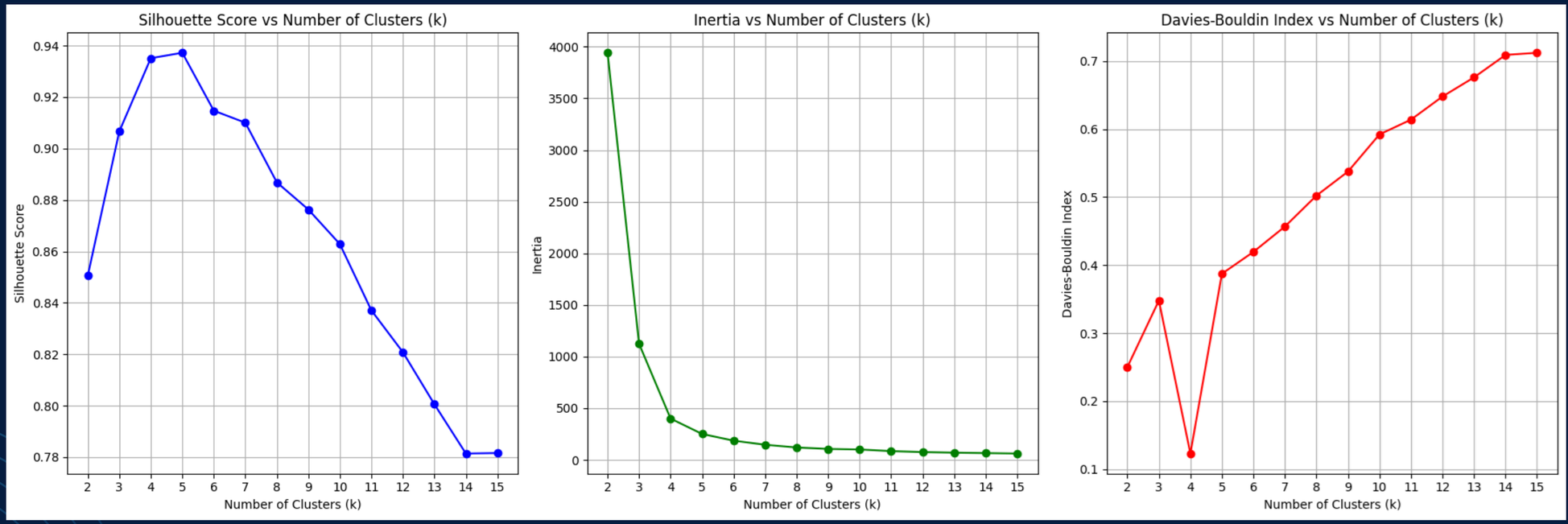
- Normalization

CLUSTERING - ELBOW METHOD

Elbow Method is used to:

- Identify the optimal number of clusters with K-Means
- Find the point where increasing k provides minimal improvement calculating performance metrics.
- Set k = 15 as the maximum number of clusters

Model	k	Silhouette	Inertia	DBI
TF-IDF	13	0.0099	28387.7595	8.6726
Glove	2	0.1188	96329.2997	2.5578
Doc2Vec	3	0.0347	5215163.6658	8.7797
BERT fine-tuned	5	0.9372	250.9741	0.3873

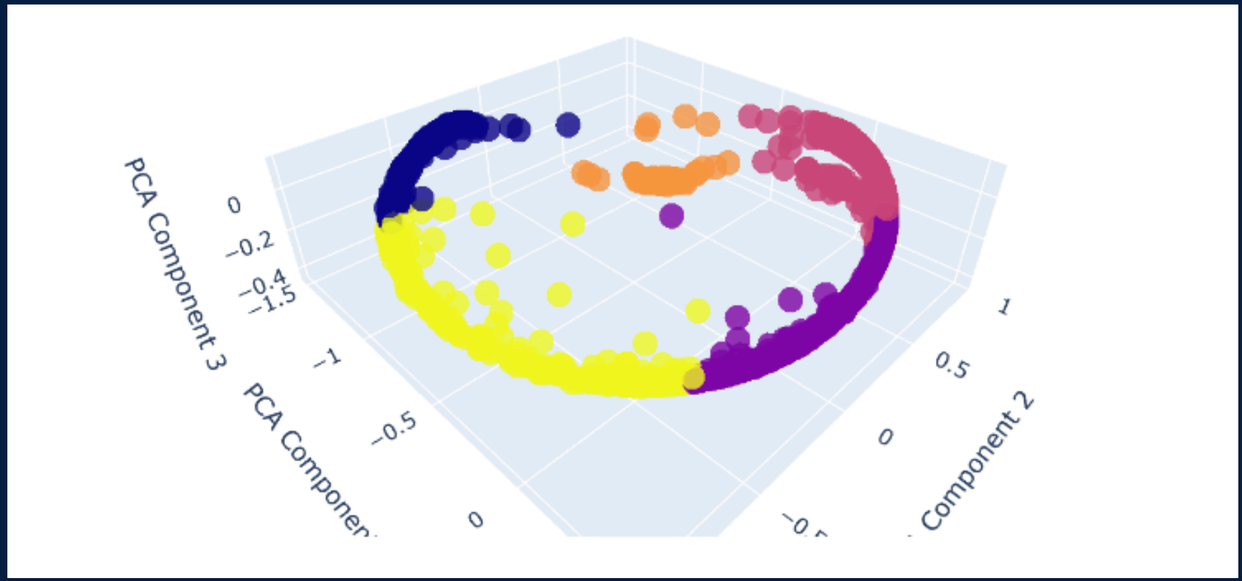


Metrics achieved with K-Means using BERT representations

RESULTS WITH K = 5

- Investigate whether the clusters created with fine tuned BERT can represent the **5 classes of stars rating associated with reviews**.
- Rand index was used to assess the correlation between the clustering results and the star ratings, which served as the ground truth.
- PCA algorithm is used to keep and then visualize the instances in 3 dimensions.

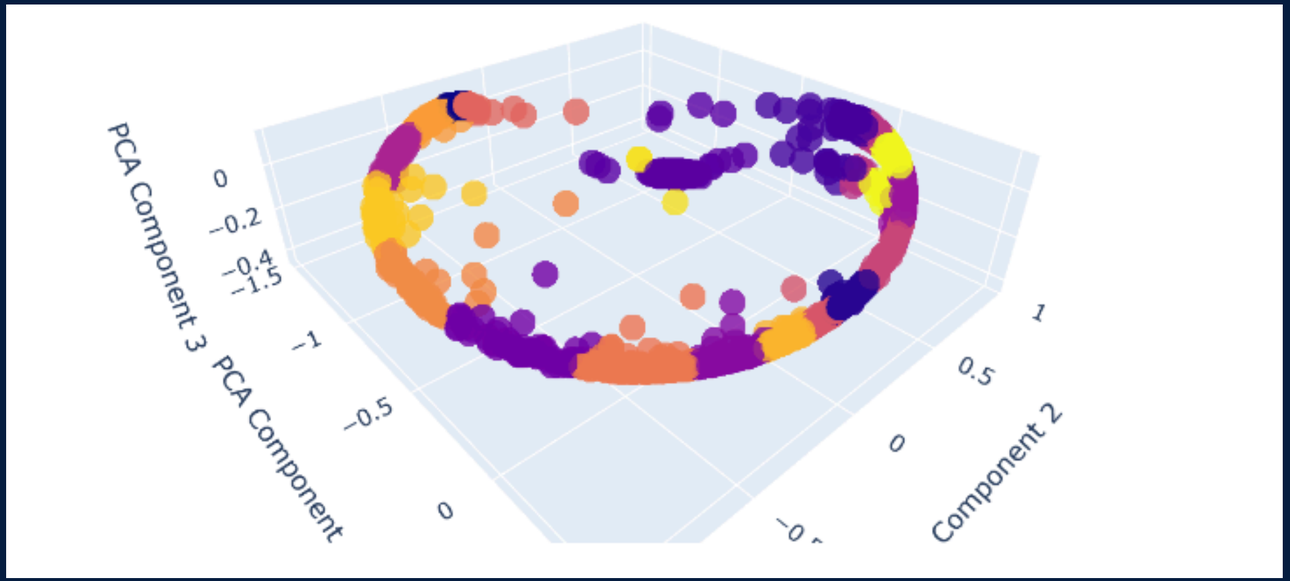
Rand Index obtained = **0.5**



RESULTS WITH K = 19

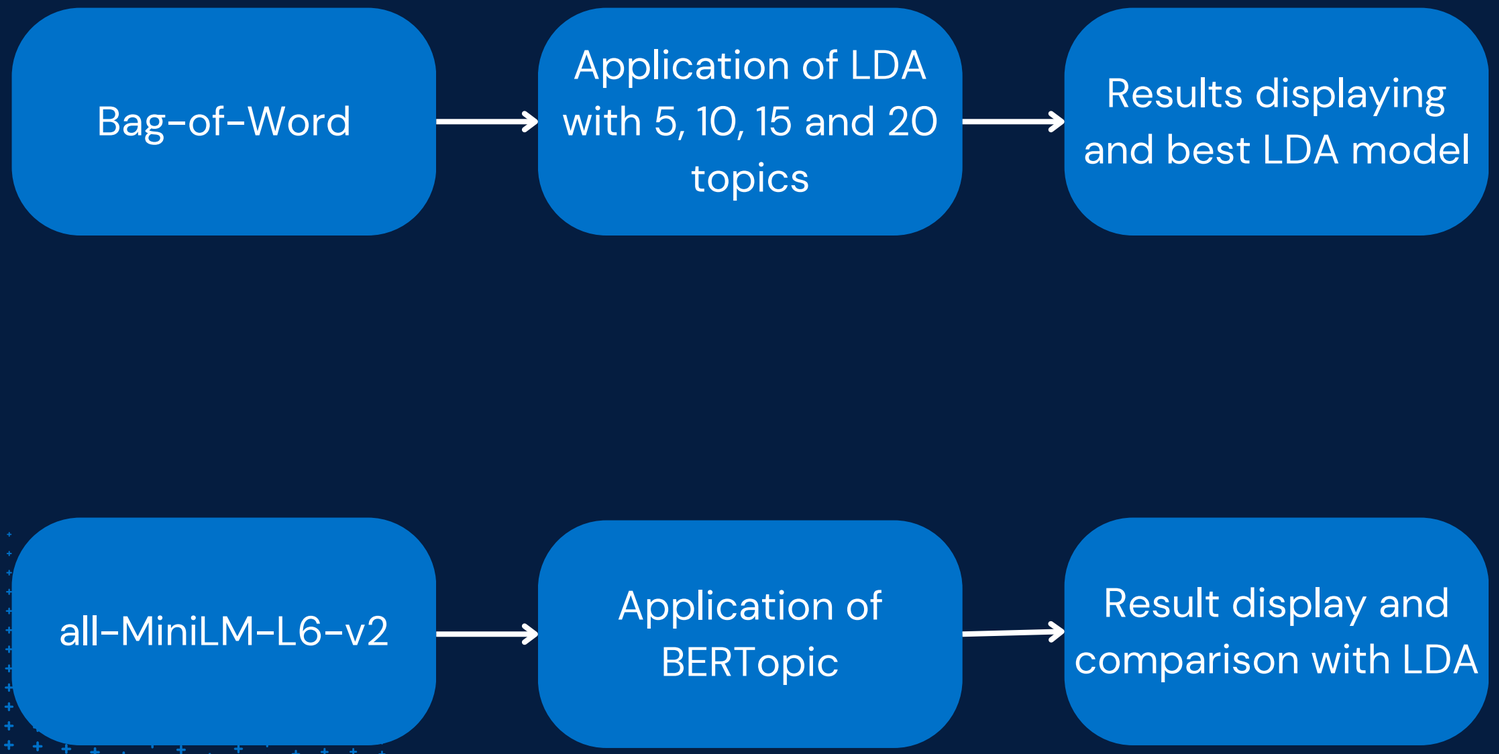
- Investigate whether the clusters created with fine tuned BERT can represent the **19 classes of review categories**.
- Rand Index was employed to evaluate the relationship between the clustering results and review categories used as ground truth.
- The PCA algorithm is applied to reduce the data to 3 dimensions, enabling the visualization of instances in a lower-dimensional space.

Rand Index obtained = **0.61**



TOPIC MODELING

Here are reported the experimental setup of the topic modeling task.



The results provided by the topic modeling algorithms are:

- evaluation based on coherence, perplexity and diversity
- Word distribution among topics
- Topic distribution among documents

RESULTS COMPARISON

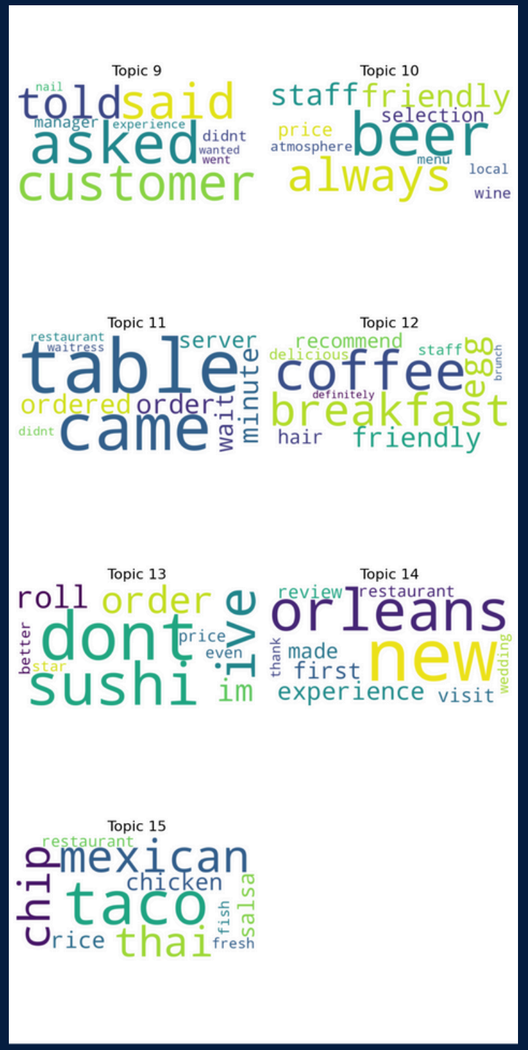
Here are reported the metric evaluations of all models to make a comparison

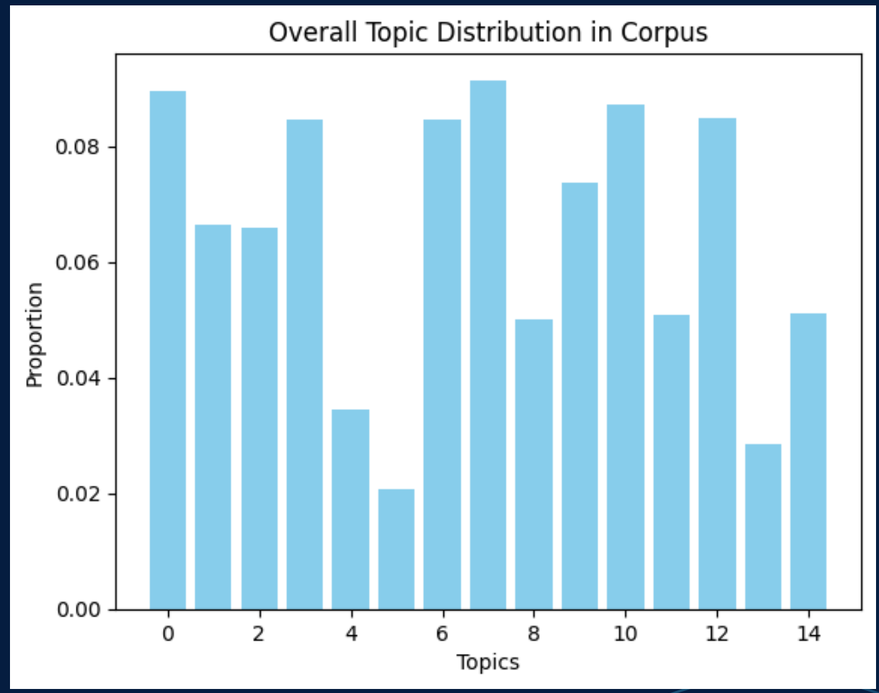
Model	Coherence	Perplexity	Diversity
LDA 5 topics	0.43	-7.9	0.77
LDA 10 topics	0.44	-8.3	0.87
LDA 15 topics	0.49	-8.9	0.88
LDA 20 topics	0.48	-9.2	0.89
BERTopic	0.63	-	0.99

Best overall model: BERTopic

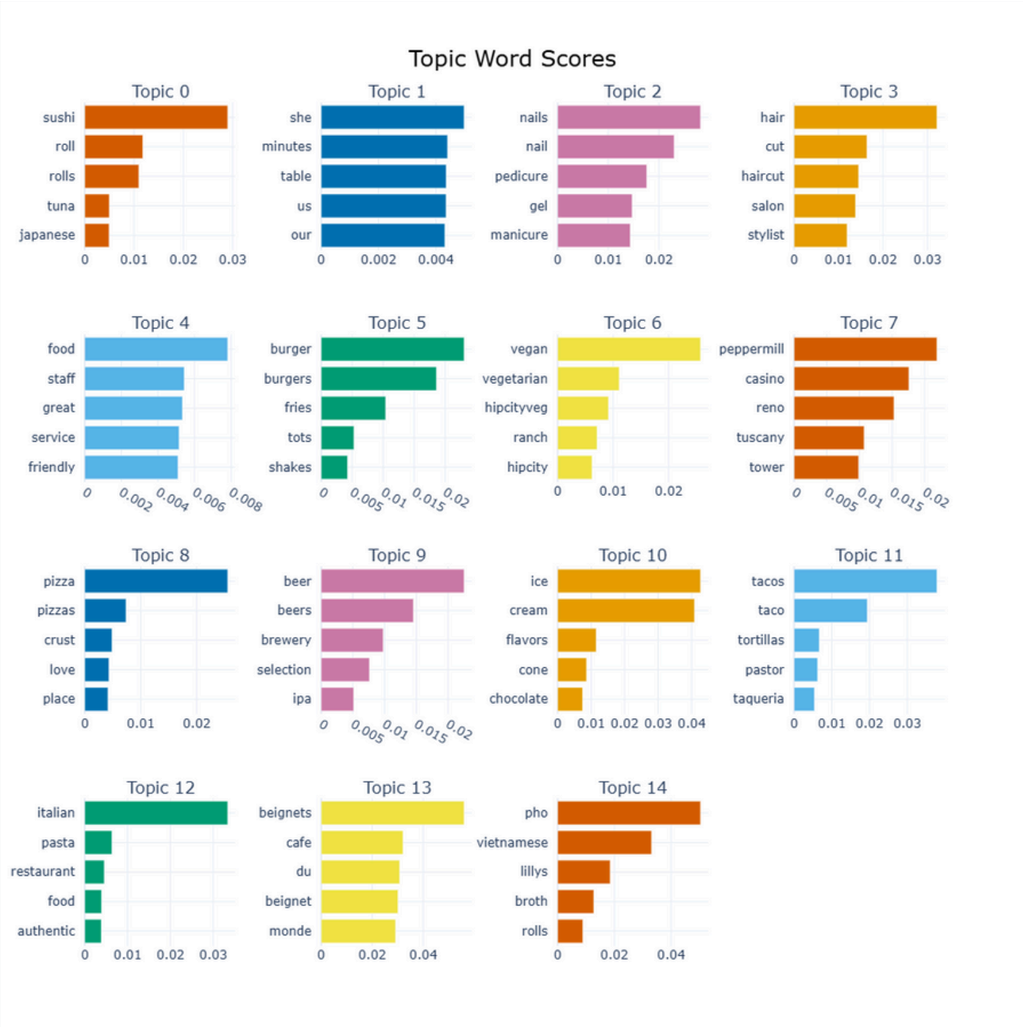
Best LDA: 15 topics

LDA 15 TOPICS RESULTS





BERTOPIC RESULTS



The best topic extraction model is BERTopic, the results have the following characteristics:

- Topics distinguish different business activities in the dataset
- Word among topics look like related
- Word among different topic look like not related

Thank's For Watching

Andrea D’Amicis 869008
Gabriele Sormani 869217
Alisia Sallemi 866453