CLASSIFICATION OF AMAZON REVIEWS BY TOPIC

Text mining and Search

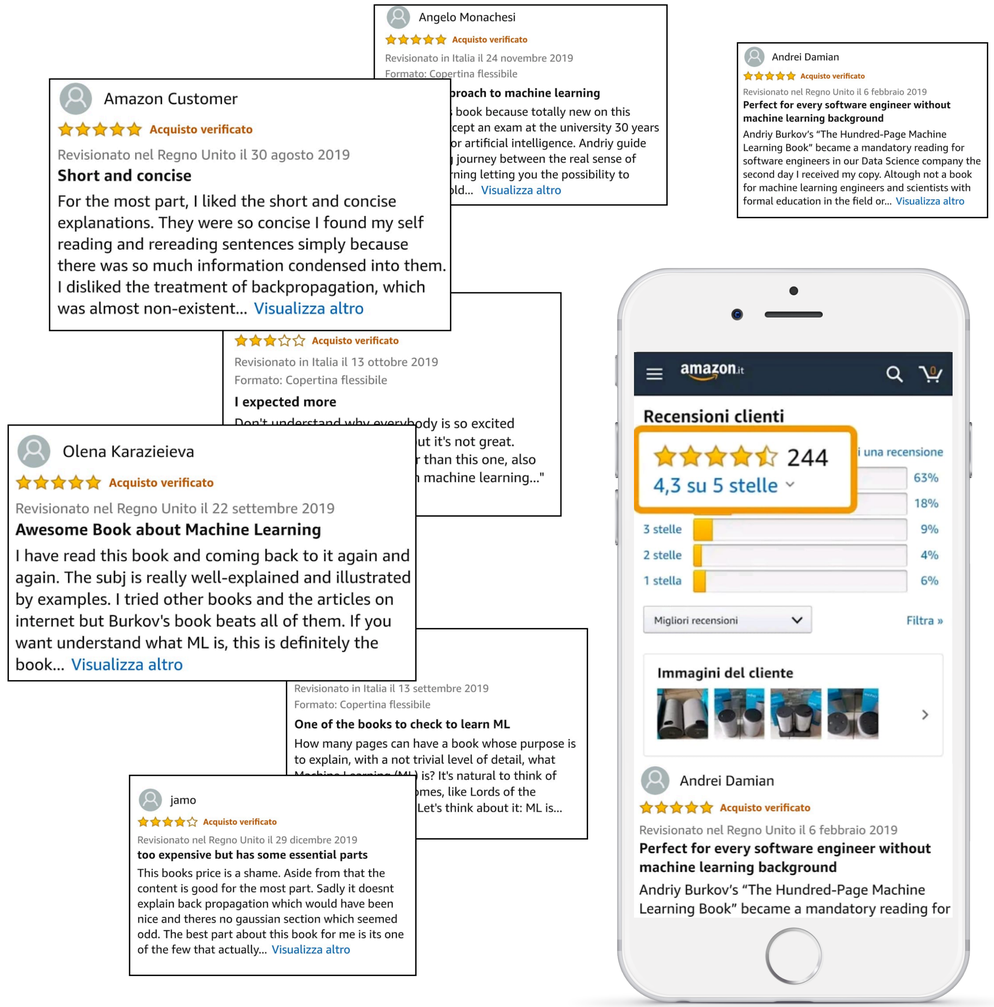
Academic Year 2019/2020

PROJECT WORKFLOW

THE TASK

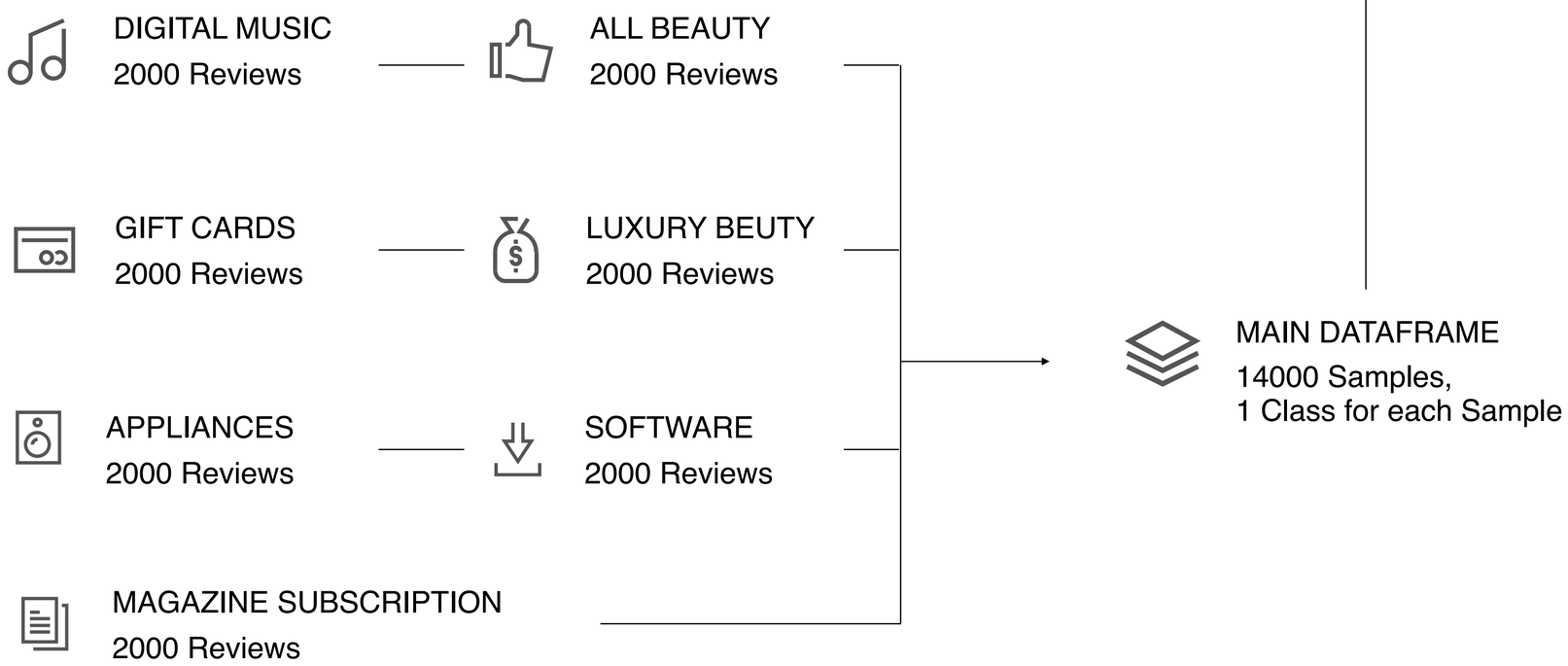
Text Classification of 14000 Amazon Reviews in 7 topics:

- DIGITAL MUSIC
  2000 Reviews
- ALL BEAUTY
  2000 Reviews
- GIFT CARDS
  2000 Reviews
- LUXURY BEUTY
  2000 Reviews
- APPLIANCES
  2000 Reviews
- SOFTWARE
  2000 Reviews
- MAGAZINE SUBSCRIPTION
  2000 Reviews



PROJECT WORKFLOW

THE PIPELINE



PREPROCESSING

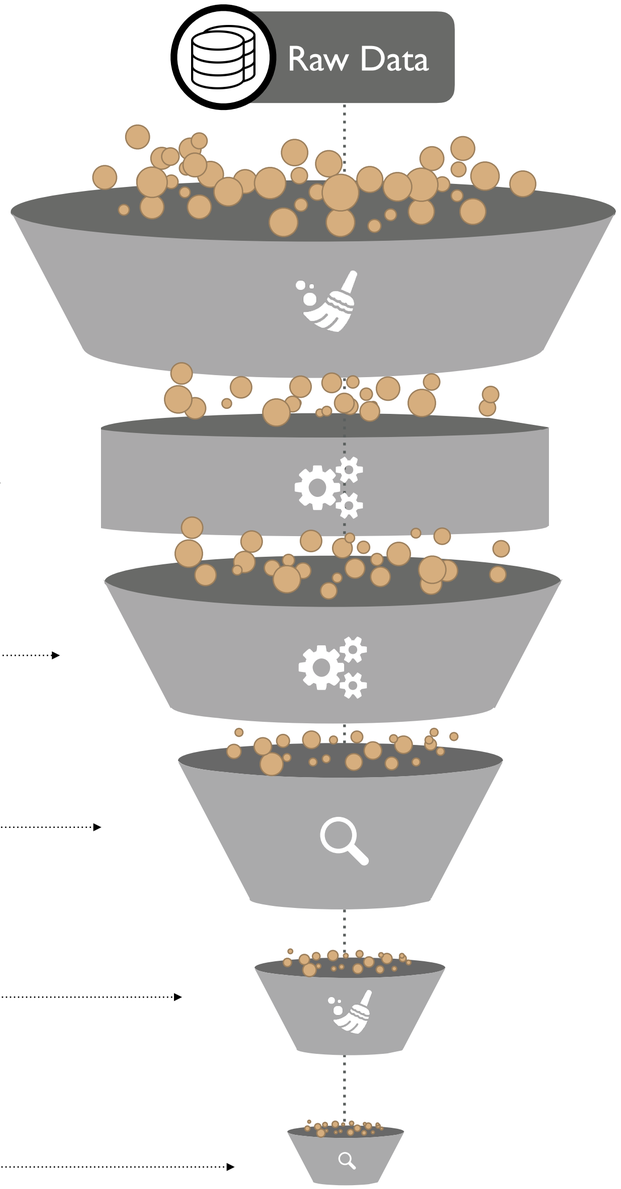
TEXT REPRESENTATION

DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

1. Removing all but characters
2. Normalization
3. Remove n-duplicate characters
4. Tokenization
5. Stopwords removal
6. Stemming



PREPROCESSING

TEXT REPRESENTATION

DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

Bag of Words

PREPROCESSING

TEXT REPRESENTATION

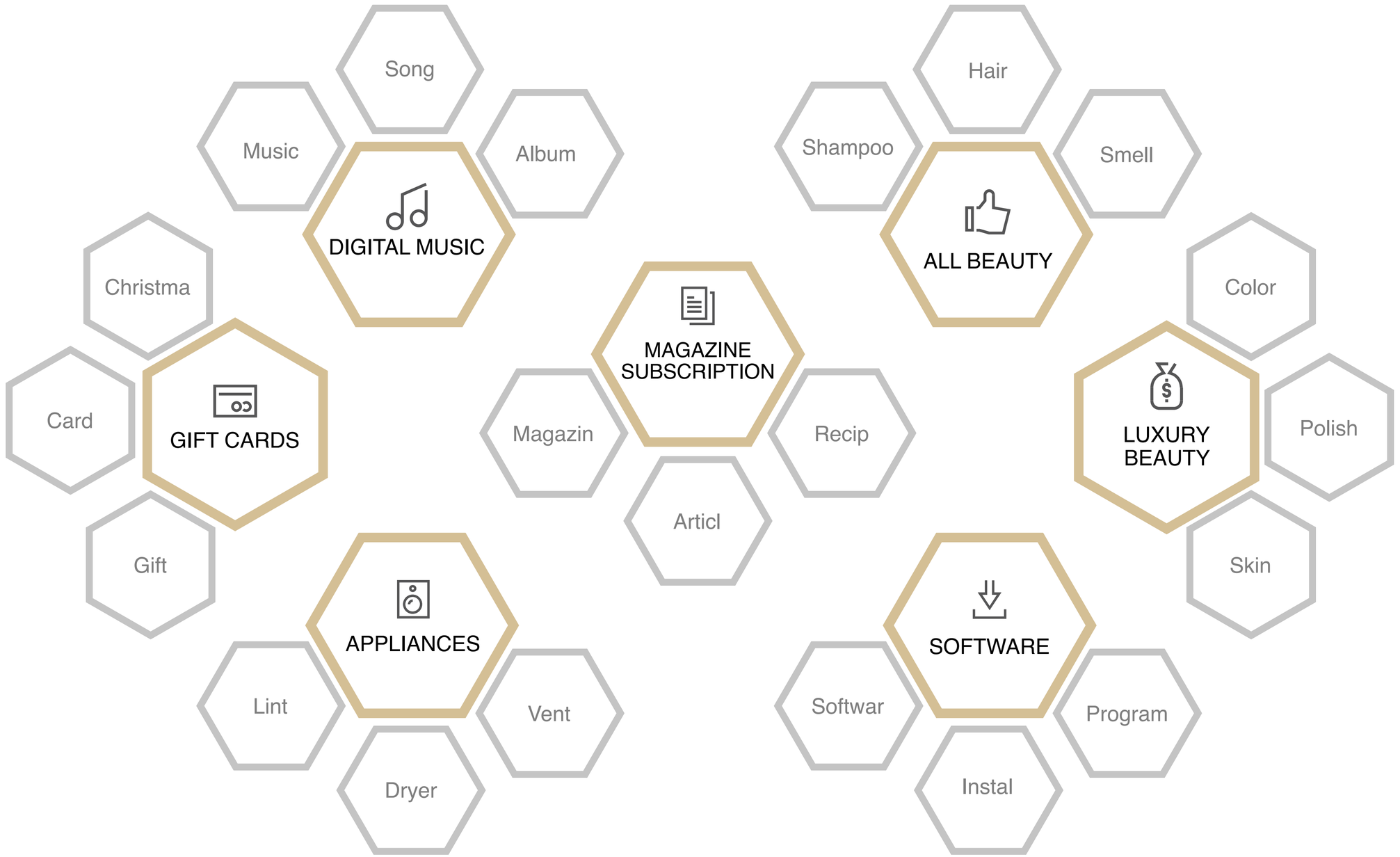
DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

Tf-Idf

Most Correlated Tokens



PREPROCESSING

TEXT REPRESENTATION

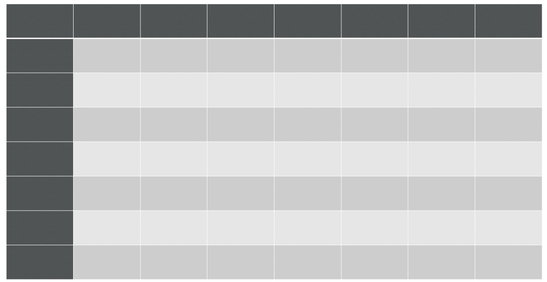
DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

Features Selection

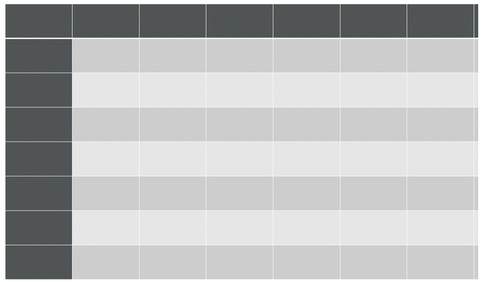
Tf-Idf Matrix



14000 x 4852

Maintaining only the best 75 % features according to χ^2 independence test

Tf-Idf Matrix Reduced



14000 x 3639

Execution Time:

1 second

PREPROCESSING

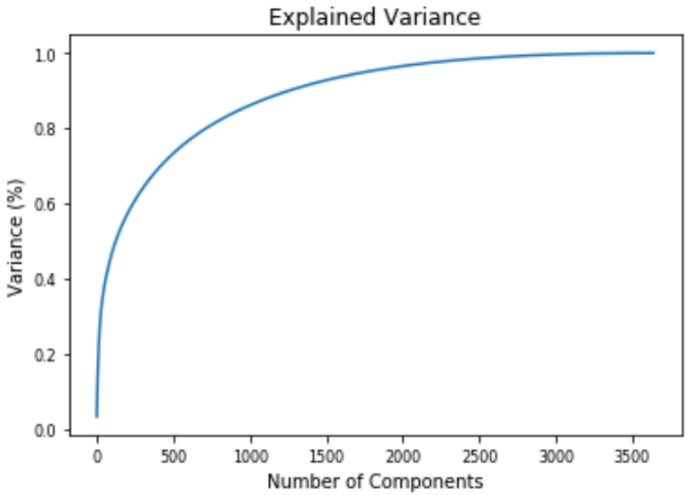
TEXT REPRESENTATION

DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

Features Synthetization



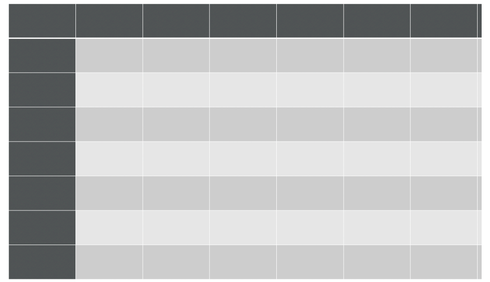
Maintaining the first 2000 components that can explain more than 90% of Variance

Execution Time:

76 seconds (fit)

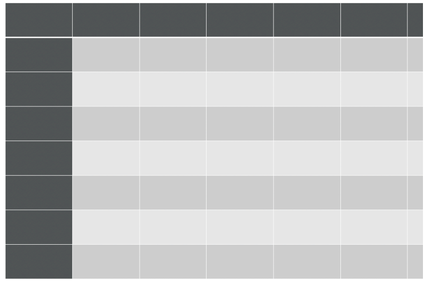
117 seconds (fit-transform)

Tf-Idf Matrix Reduced



14000 x 3639

PCA Matrix



14000 x 2000

PREPROCESSING

TEXT REPRESENTATION

DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

5-fold Cross Validation

Training Set: 12600 examples (90%)

Test Set: 1400 examples (10%)

1 **SVC**

Support Vector Classificator

Penalty: Norm L2

Loss: Squared Hinge

C: 1.0

Training Time: 46 seconds

Training Costs: 239 seconds

2 **LR**

Logistic Regression

Penalty: Norm L2

Solver: lbfgs

C: 1.0

Training Time: 65 seconds

Training Costs: 258 seconds

3 **MLP**

MultiLayer Perceptron

HL sizes: 100,50,50,50

Activation: ReLU

Solver: AdaM

Training Time: 336 seconds

Training Costs: 529 seconds

PREPROCESSING

TEXT REPRESENTATION

DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

5-fold Cross Validation

Training Set: 12600 examples (90%)

Test Set: 1400 examples (10%)

1 **SVC**
Support Vector Classificator

Mean Accuracy:
0.865397

Standard Deviation:
0.004690

2 **LR**
Logistic Regression

Mean Accuracy:
0.859206

Standard Deviation:
0.008415

3 **MLP**
MultiLayer Perceptron

Mean Accuracy:
0.833651

Standard Deviation:
0.012083

PREPROCESSING

TEXT REPRESENTATION

DIMENSIONALITY REDUCTION

CLASSIFICATION

EVALUATION

Test set predictions

Training Set: 12600 examples (90%)

Test Set: 1400 examples (10%)

1

SVC

Support Vector Classificator

Mean Accuracy:

0.865397

Standard Deviation:

0.004690

	precision	recall	f1-score	support
Digital_Music	0.81	0.79	0.80	199
All_Beauty	1.00	0.97	0.98	183
Software	0.86	0.87	0.86	202
Gift_Cards	0.75	0.89	0.81	210
Luxury_Beauty	0.84	0.82	0.83	204
Appliances	0.91	0.88	0.89	214
Magazine_sub	0.97	0.88	0.92	188
accuracy			0.87	1400
macro avg	0.88	0.87	0.87	1400
weighted avg	0.87	0.87	0.87	1400

Classification Costs:

1 second

CLASSIFICATION OF AMAZON REVIEWS BY TOPIC

CONCLUSIONS

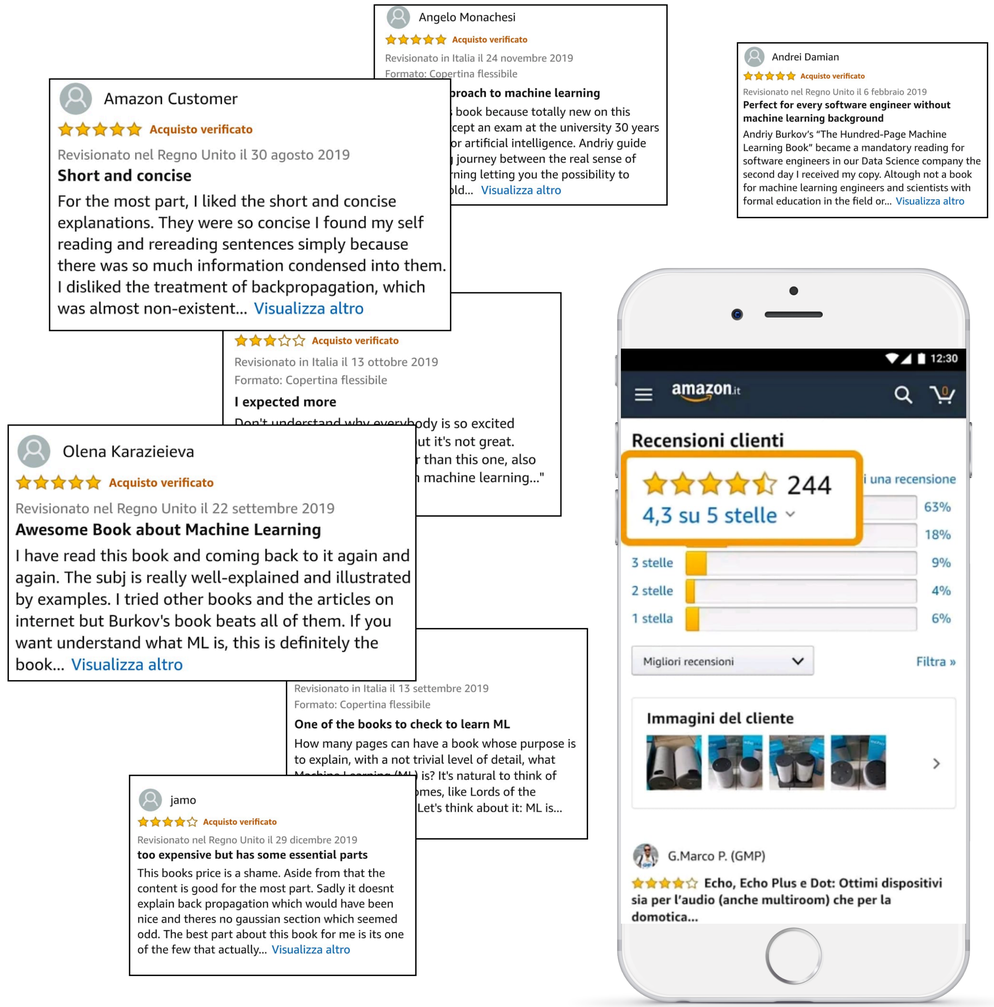
- ML algorithms are able to correctly classify Amazon Reviews by topic.

ISSUES

- High Computational Costs in features synthetization;
- Limited number of observation for each category due to prohibitive computational costs.

NEXT STEPS

- Scale the problem in dimension with less expensive dimensionality techniques, such as LDA.



THANKS FOR YOUR ATTENTION

Raffaele Anselmo - 846842

Lorenzo Pastore - 847212