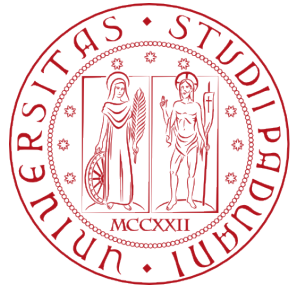
Sentiment Analysis on Online Automotive Forums

Candidate: Giuseppe Ravagnani
Supervisor: Prof. Nicola Ferro
Co-Supervisor: Davide Lappon

September 30, 2019



Università degli Studi di Padova

Introduction

Comparison with Twitter

Example block

@united I do not see where it talks about military baggage fees. Can you please guide me. Thanks #usairline

Example block

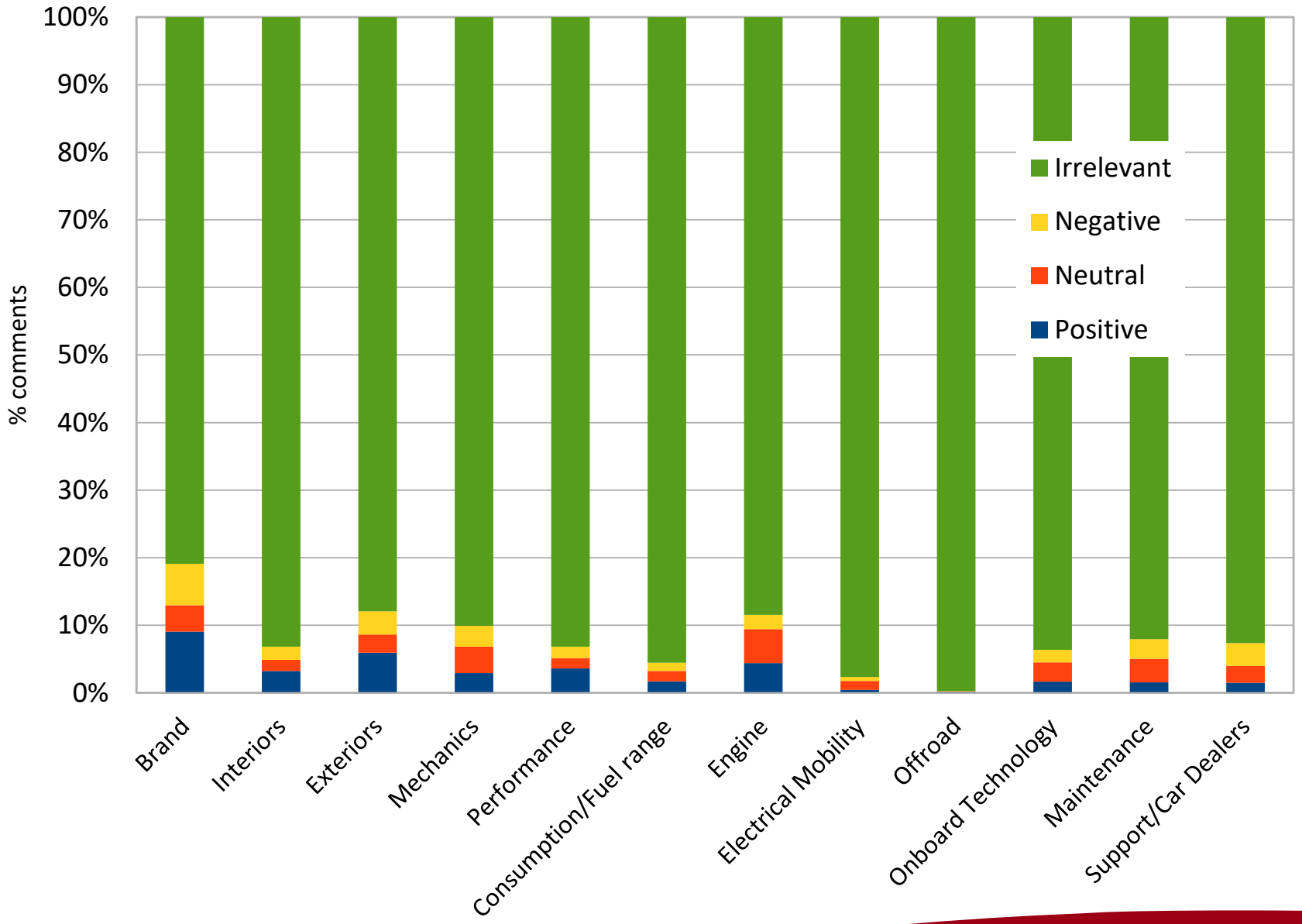
Sono reali calcolati nel arco del tutto anno nel estate qualcosa in più causa gomme di 17" e climatizzatore nel inverno un po di meno. Per quanto riguarda le autostrade quelle che percorro io principalmente la A4 e molto congestionata cosi spesso la media e 110-115 km/h che ovviamente influisce positivamente a i consumi. Ma quello che mi piace di più è assenza dei guasti. Sulla vecchia Accord il primo guasto lo ho avuto a 200000 km si è rotto il termostato della clima. Ogni tanto faccio giro di altri forum e leggo delle turbine rotte catene di distribuzione progettate male iniettori fatti male mah nel 2015 per me sono le cose incomprensibili. Con tutti gli difetti che può avere preferisco la Honda.

Dataset Gathering

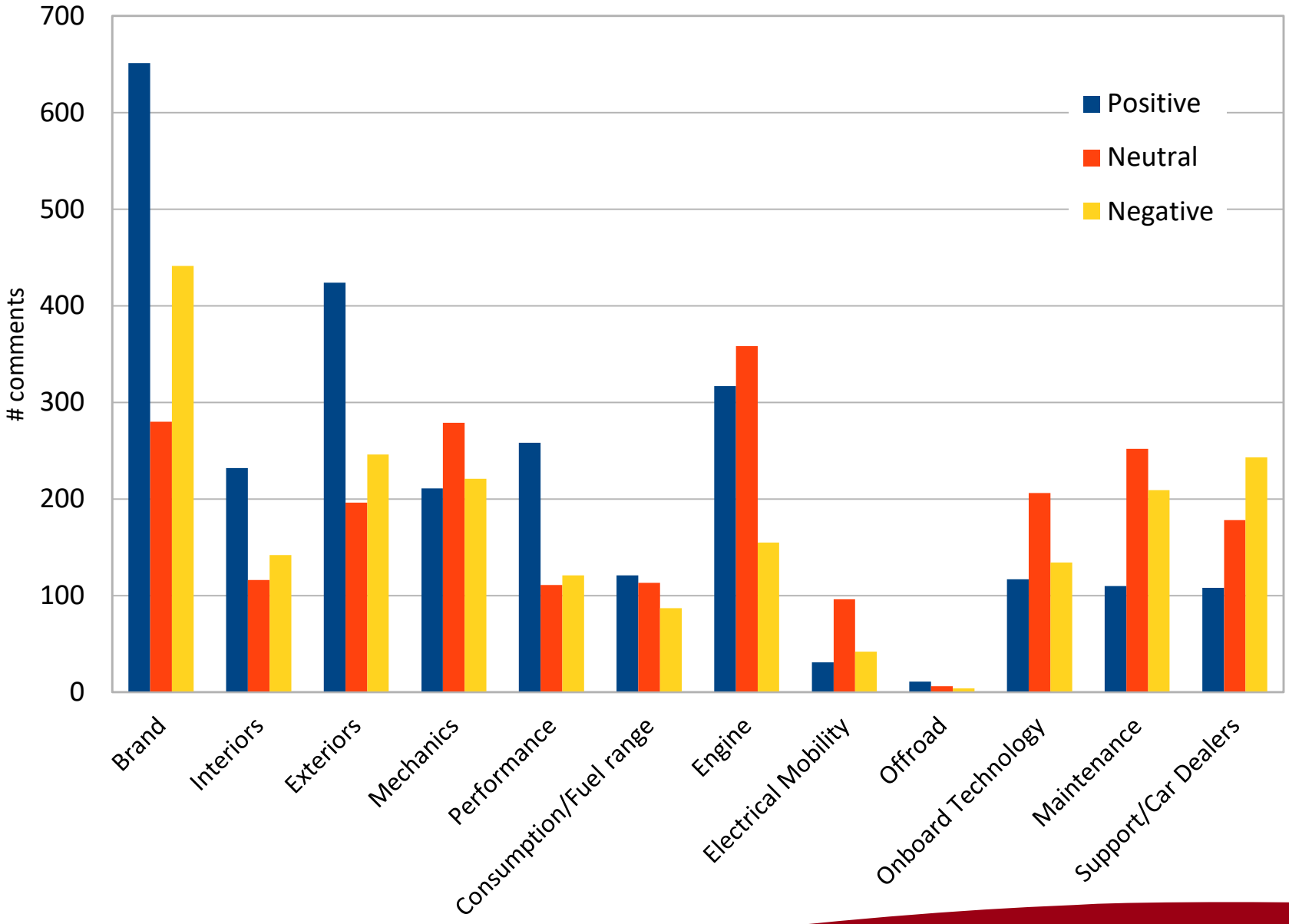
The creation of a suitable dataset is mandatory to train machine learning algorithm.

- Crawled some of the most visited Italian automotive forums (Quattroruote, Autopareri, Bmwpassion, HDmotori, Porschemania, Forumelettrico)
- Comments have been annotated with respect to the classes "Brand", "Interiors", "Exteriors", "Mechanics", "Performance", "Consumption", "Engine", "Electrical mobility", "Off-road" and "Technology", picking the labels "positive", "negative", "neutral" or "irrelevant".
- From a total of 1,200,000 crawled comments, 7,183 have been manually annotated

Dataset Statistics



Dataset Statistics



Baseline algorithm

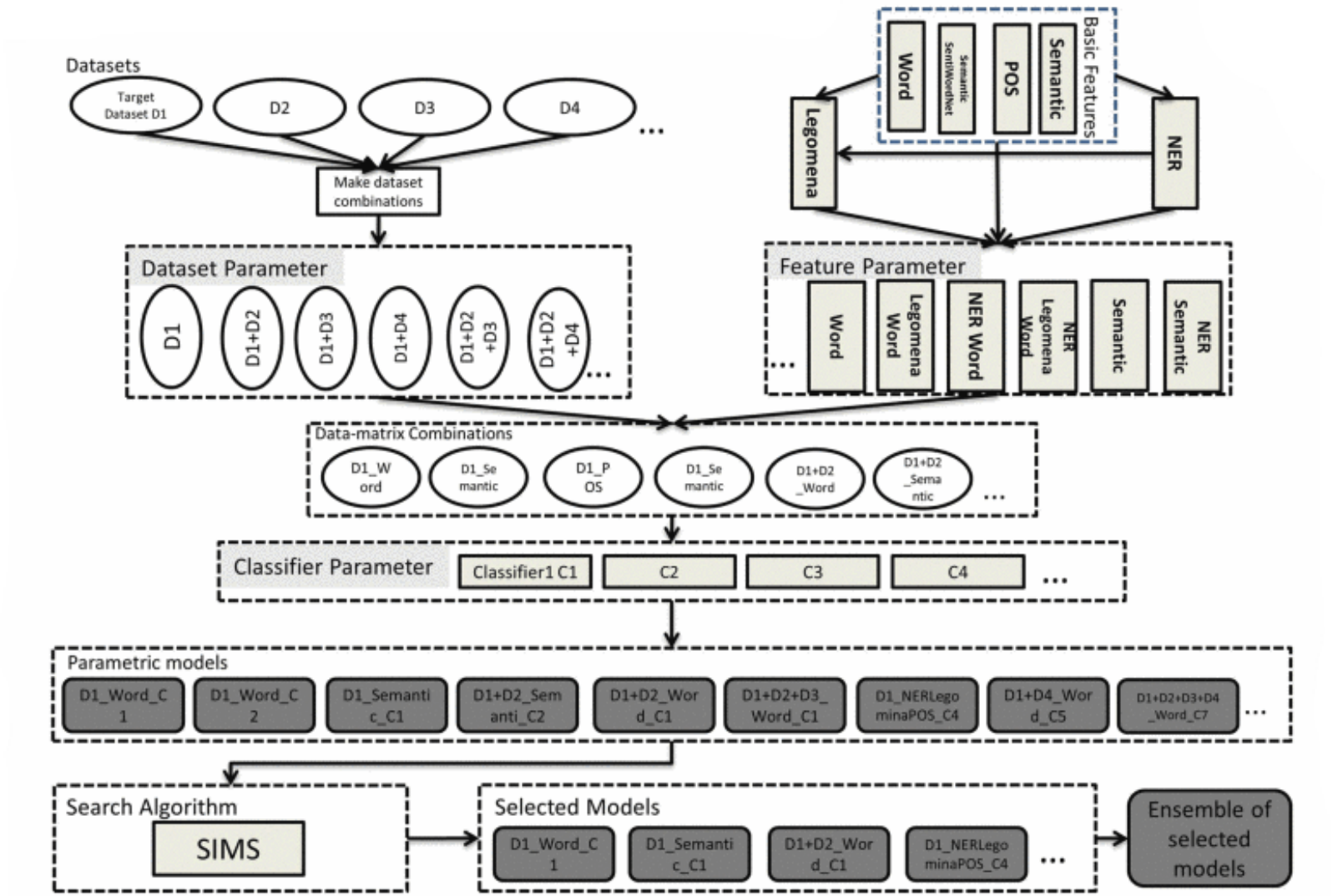
→ Our dataset preprocessing:

1. encoding correction
2. lowercase, punctuation and stopwords removal
3. "http://someurl" → "URL"
4. replacing domain-specific tokens with common string (distances, speed, consumption, weight, power, ...)
5. stemming

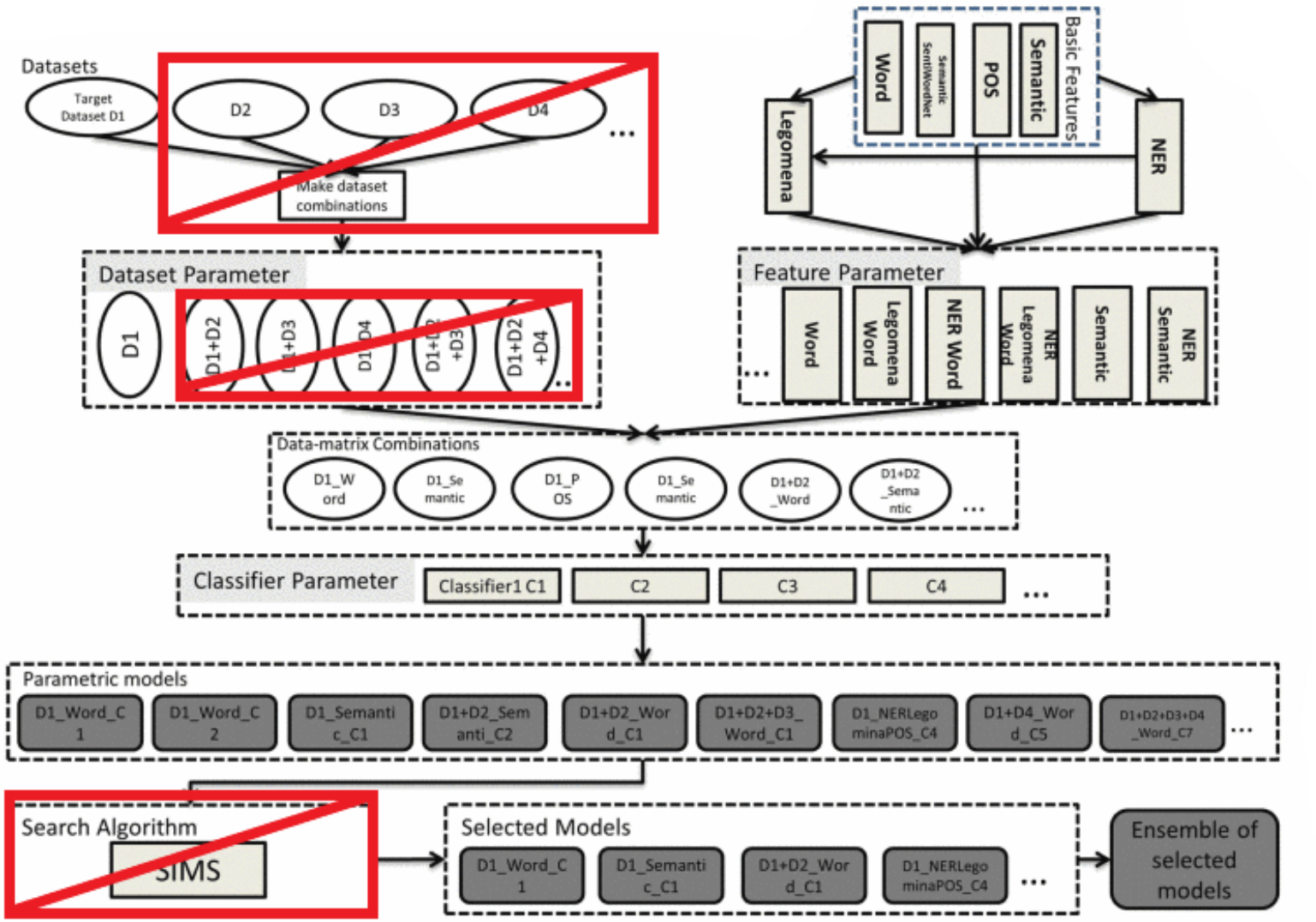
→ Support Vector Machine (SVM) Classifier with Term Frequency - Inverse Document Frequency (TF-IDF) features vectorization

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|}, \quad idf_i = \log \frac{|D|}{|\{d : i \in d\}|}, \quad tf\text{-}idf_{i,j} = tf_{i,j} \times idf_i$$

Bootstrap Ensemble Framework (BPEF)

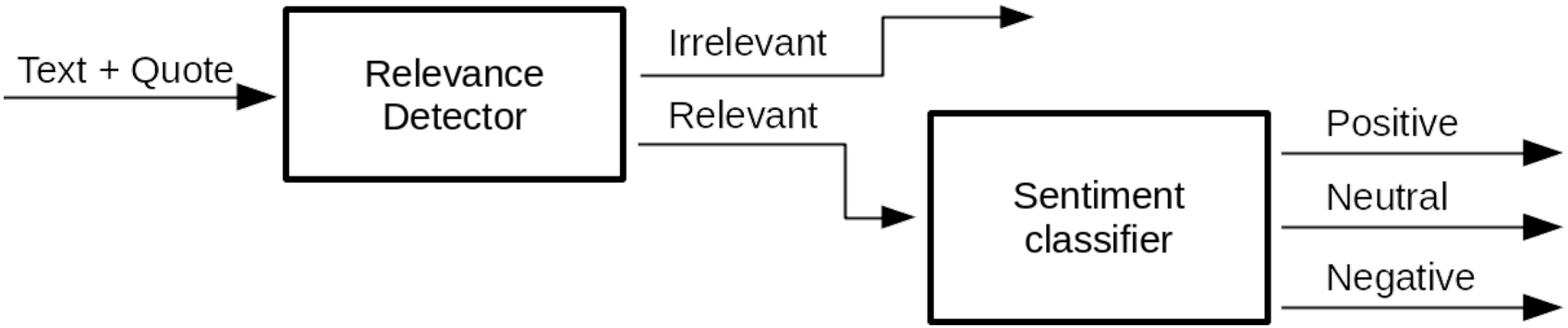


Bootstrap Ensemble Framework (BPEF)



Cascade Classifier

Implementation of a cascade classifier for the four-label classification.

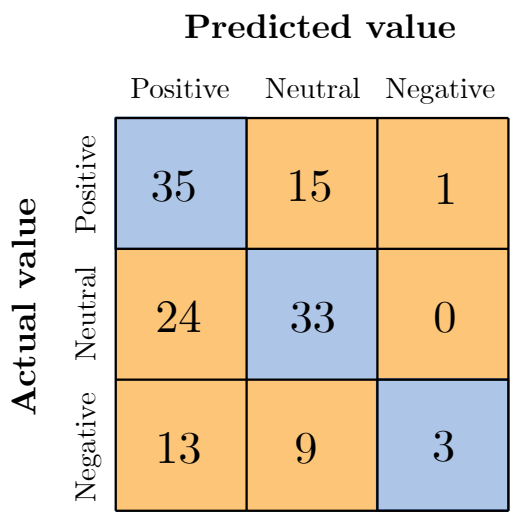


- Logistic Regression relevance detector
- BPEF Sentiment classifier

Results: Tests on our Dataset

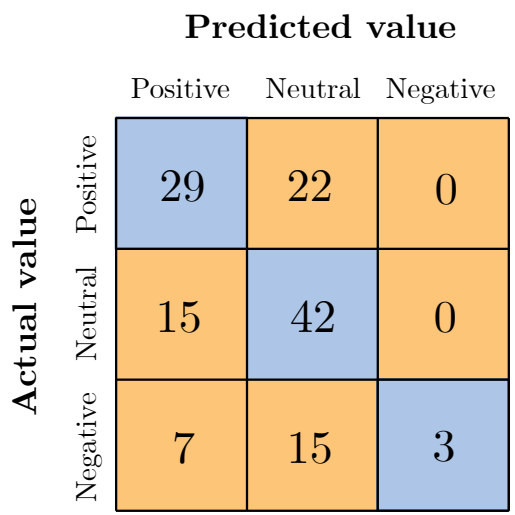
Sentiment classification

SVM



F1-macro	0.451

BPEF



F1-macro	0.467

Results: Tests on our Dataset

4-labels classification

SVM

Predicted value

Actual value	Irrelevant	Positive	Neutral	Negative
Irrelevant	1000	11	6	0
Positive	32	15	4	0
Neutral	38	10	9	0
Negative	20	3	2	0

F1-macro	0.378

Cascade classifier

Predicted value

Actual value	Irrelevant	Positive	Neutral	Negative
Irrelevant	933	18	61	5
Positive	19	23	9	0
Neutral	20	4	33	0
Negative	11	3	3	8

F1-macro	0.556

Data Visualization

dataset_data_visualization - Panamera

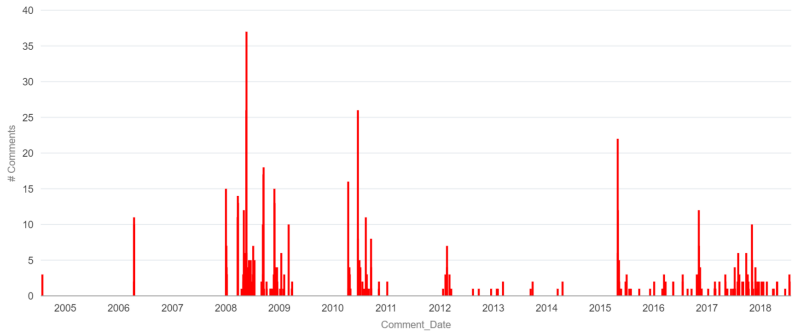
Filtro espressione

PANAMERA COMMENTS

839

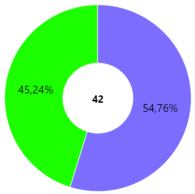
PANAMERA COMMENTS FLOW

Filtro espressione



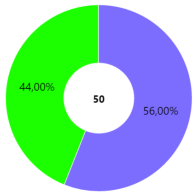
ENGINE SENTIMENT PANAMERA BEFORE 2015

Filtro espressione



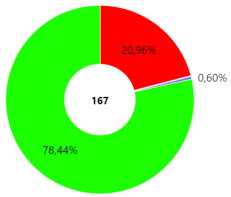
ENGINE SENTIMENT PANAMERA AFTER 2015

Filtro espressione



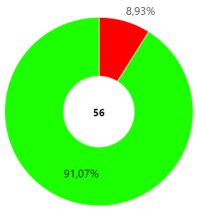
EXTERIORS SENTIMENT PANAMERA BEFORE 2015

Filtro espressione



EXTERIORS SENTIMENT PANAMERA AFTER 2015

Filtro espressione



Conclusions

- BPEF model overcomes baseline approach
- Implemented model can be considered reliable for sentiment classification for Italian automotive forums
- Expanding the dataset, same algorithms should improve their scores
- Design a more sophisticated relevance detector for identifying the topic
- Integration in a production system: scheduled crawler, database and business intelligence software