Sentiment Analysis on Online Automotive Forums

Candidate: Giuseppe Ravagnani
Supervisor: Prof. Nicola Ferro

September 30, 2019

Università degli Studi di Padova

Outline

1. Introduction
2. Sentiment Analysis Algorithms
3. Dataset
4. My Implementation
5. Results
6. Data Visualization
7. Conclusions & Enhancements

Introduction (1)

Sentiment Analysis is an active area of research in Natural Language Processing, motivated to improve the automated recognition of sentiment expressed in text

- From the whole scenic of online social media, Twitter is the most used to achieve contents
- Most existing techniques for sentiment classification involve supervised learning
- Used for collecting people's sentiment information about a given topic

Introduction (2)

→ **Dataset**: the need of a set of already annotated data for applying machine learning models

→ **Development**:

- Twitter Sentiment Analysis Algorithm
- Relevance Detection for class "Engine"
- Sentiment Classification for class "Engine"
- Cascade Classifier

→ **Goal**: Develop an aspect-based sentiment analysis tool in automotive field for a specific brand

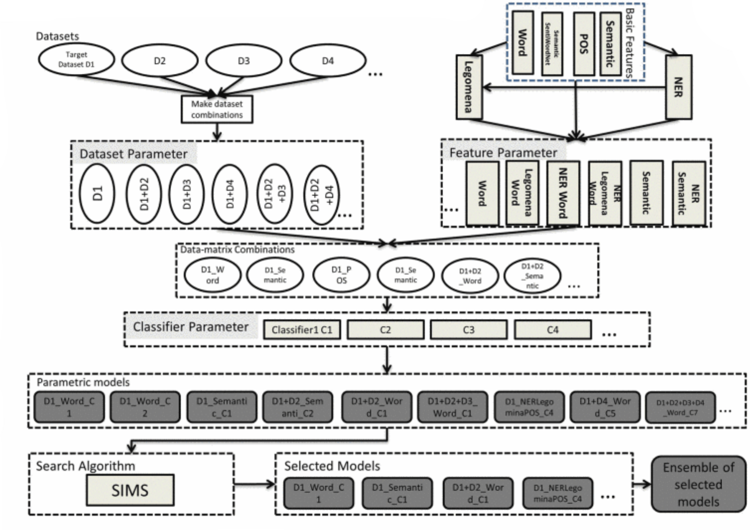
→ **Collaboration**: Reply Technology for commission of Porsche

State of the Art Algorithms

From: Zimbra, David & Abbasi, Ahmed & Zeng, Daniel. (2018).
The State-of-the-Art in Twitter Sentiment Analysis:
A Review and Benchmark Evaluation.

	Average	Pharma	Retail	Security	Tech	Telco	Ensemble
BPEF	71.38	67.81	65.24	75.32	76.30	72.21	yes
NRC	71.33	75.26	64.93	76.39	64.96	75.08	no
Webis	71.41	76.16	64.40	77.37	63.68	75.46	yes

Bootstrap Ensemble Framework (BPEF)

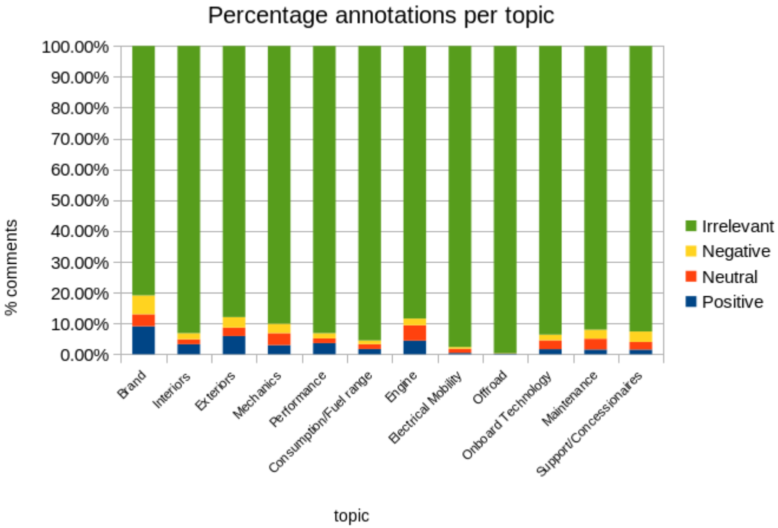


Dataset Gathering

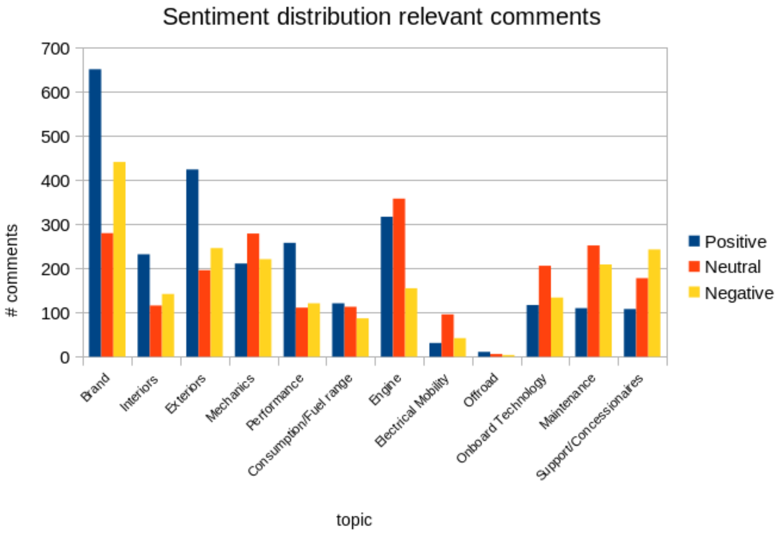
The creation of a suitable dataset is mandatory to train machine learning algorithm.

- 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



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.

Baseline Approach (2)

→ Twitter preprocessing:

1. lowercase, punctuation and stopwords removal
2. "http://someurl" → "URL"
3. "#hashtag" → "hashtag"
4. happy emoticons → "EMO_POS"
   sad emoticons → "EMO_NEG"
5. stemming

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

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

Baseline Approach (2)

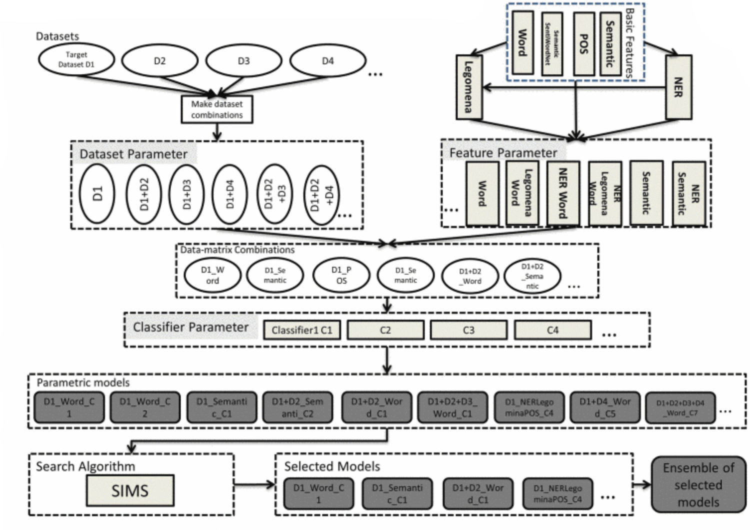
→ Our dataset preprocessing:

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

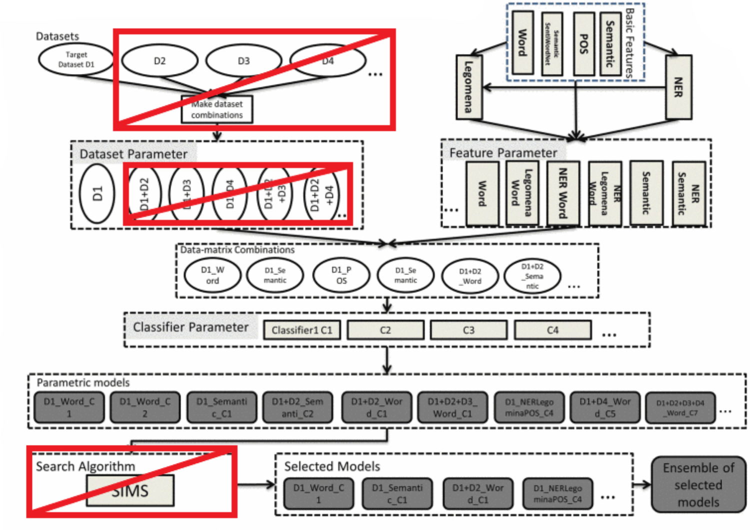
→ Support Vector Machine (SVM) Classifier with Term Frequency - Inverse Document Frequency (TF-IDF) features vectorization

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

My BPEF Implementation

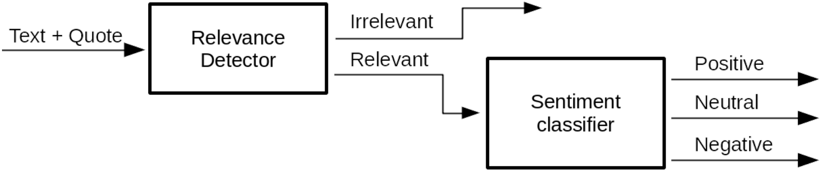


My BPEF Implementation



Cascade Classifier

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



- Logistic Regression relevance detector
- BPEF Sentiment classifier

Results: Tests on Twitter

Baseline

		Predicted value		
		Positive	Neutral	Negative
Actual value	Positive	235	44	21
	Neutral	36	196	68
	Negative	27	42	231

F-macro	0.735
Accuracy	0.736

BPEF

		Predicted value		
		Positive	Neutral	Negative
Actual value	Positive	230	55	15
	Neutral	39	221	40
	Negative	20	72	208

F-macro	0.734
Accuracy	0.732

Results: Tests on our Dataset

Relevance detection

SVM

		Predicted value	
		Irrelevant	Relevant
Actual value	Irrelevant	962	55
	Relevant	60	73

F1-macro	0.559
Recall	0.570
Precision	0.549

Logistic Regression

		Predicted value	
		Irrelevant	Relevant
Actual value	Irrelevant	933	84
	Relevant	50	83

F1-macro	0.553
Recall	0.624
Precision	0.497

Results: Tests on our Dataset

Sentiment classification

SVM

		Predicted value		
		Positive	Neutral	Negative
Actual value	Positive	35	15	1
	Neutral	24	33	0
	Negative	13	9	3

F1-macro	0.451

BPEF

		Predicted value		
		Positive	Neutral	Negative
Actual value	Positive	29	22	0
	Neutral	15	42	0
	Negative	7	15	3

F1-macro	0.467

Results: Tests on our Dataset

4-labels classification

SVM

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

F1-macro	0.378

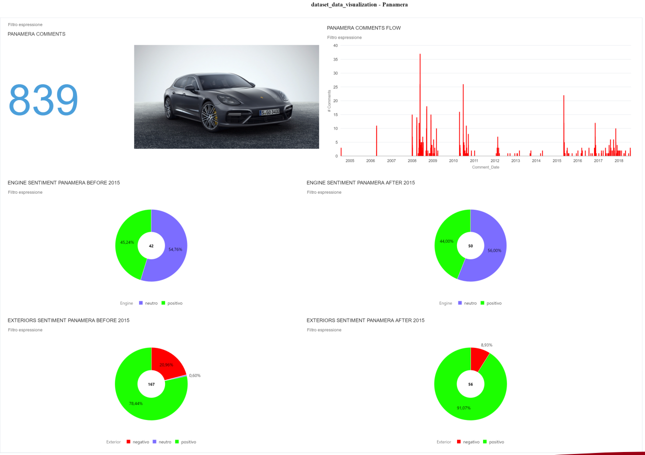
Cascade classifier

		Predicted value			
		Irrelevant	Positive	Neutral	Negative
Actual value	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

New data involving some other information have been crawled for test a use case of the classifier.



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 crawled, database and business intelligence software