

Retail Reviews Text Mining

Iper Brembate & Competitors

Master in BI & Big Data Analytics (2022/2023)

Web Data Analytics – Team project work



Team:
Michele Iannotta
Luca Izzo
Fabio Jr Lorenzini
Riccardo Licciardello

Milano, November 2023

Context

“Iper, La grande i” is a hypermarket located in 4 Italian regions: 22 stores with more than 1 million customers in a year.

The project work integrates the analysis of the customer experience to explain the economic performance of a store.



Business perspective

Allow stakeholders to:

- Integrate customer reviews from Google Maps into internal know-how.
- Identify strengths points.
- Identify areas of improvement compared to competitors.



Current Challenge & Goals

PROBLEM

Iper Brembate revenues are declining YoY compared to other Iper stores

ONGOING ACTIVITIES

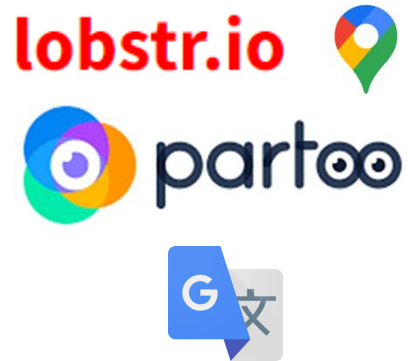
- Category analysis to identify possible issues
- Mystery shopping of customer experience with a Market Research company

PROJECT GOAL

Collect and analyze customer reviews on online channels (e.g. Google Maps) through Text Mining techniques

Process & Technologies

SCRAPING + TRANSLATION



- Lobstr.io to collect **Google Maps** reviews of 6 retail stores. **Partoo** for lper data
- More than **8k** reviews extracted and translated in English

STORAGE



Google Drive group shared folder

TEXT MINING



Orange Data Mining:

- Data Retrieval
- Preprocessing
- Word Cloud
- Sentiment Analysis
- Clustering

R Studio:

- Sentiment for most frequent words






DATA VIZ



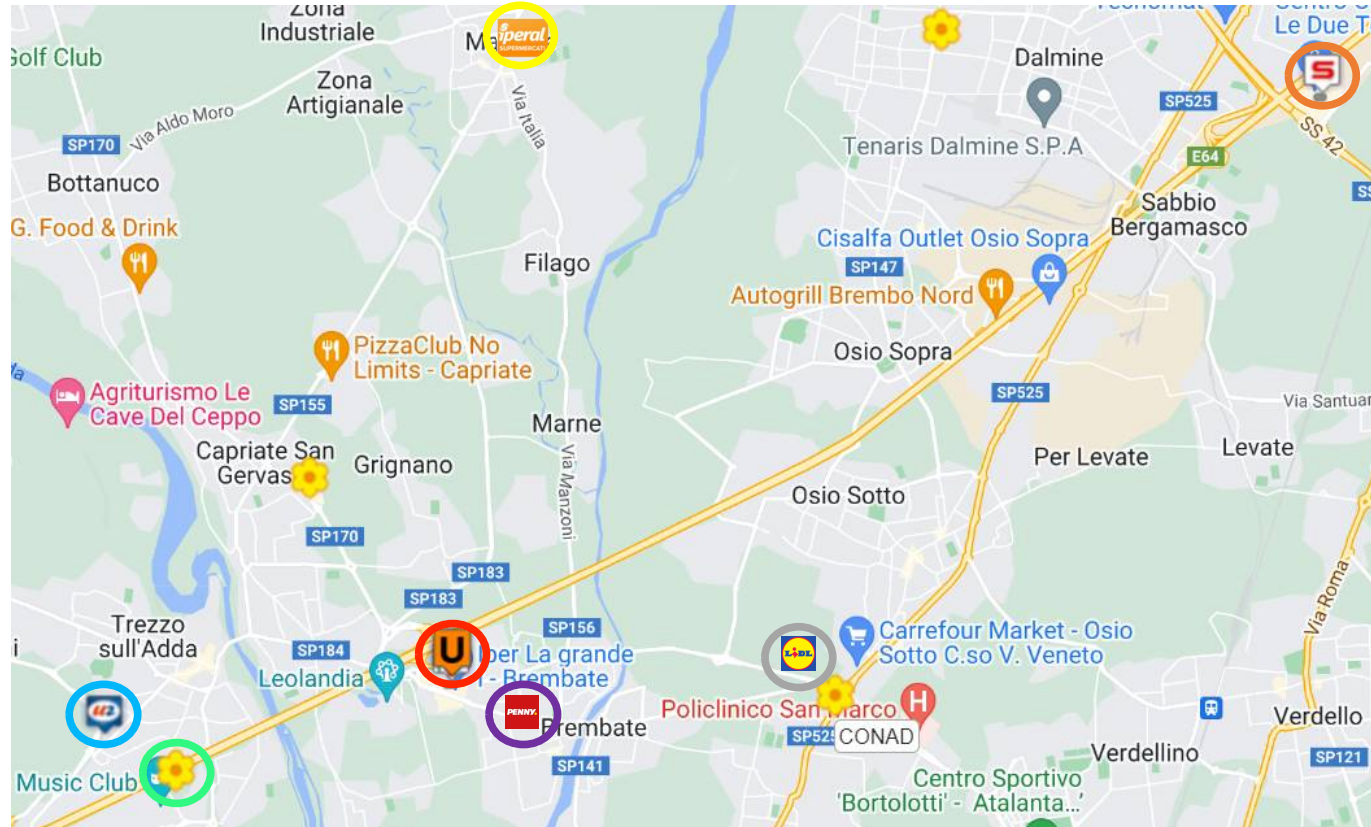
- Dashboard on PowerBI
- Create plot with Excel for most easy graphs

Rating Reviews Google Brembate vs. 22 Iper stores

Analyzing the rating distribution based on Google Maps, Iper Brembate has more negative scores than the average of overall Iper stores.

GOOGLE MAPS RATING SCORE	IPER BREMBATE	AVERAGE 22 STORES IPER	Δ BREMBATE VS 22 IPER STORES
	9%	5%	+ 4%
	6%	3%	+ 3%
	15%	9%	+ 6%
	30%	28%	+ 2%
	40%	55%	- 15%

Rating Reviews Iper Brembate vs. Competitors



GOOGLE MAPS RATING SCORE

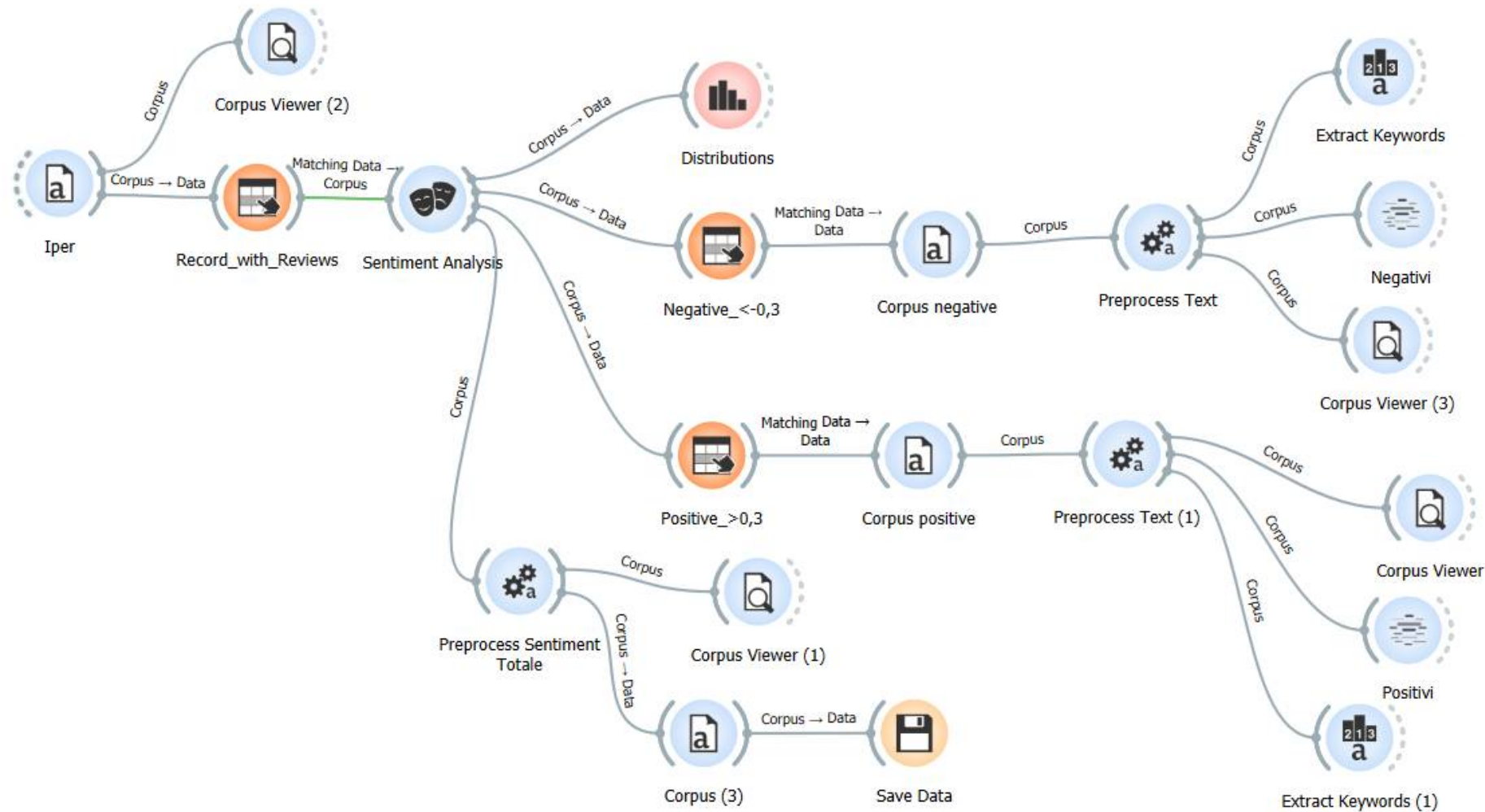
● Iper, Brembate	3,9
● Esselunga, Stezzano	4,3
● Conad, Trezzo Sull'Adda	4,2
● Iperal, Presezzo	4,3
● Unes, Trezzo sull'Adda	4,1
● Penny, Brembate	4,0
● Lidl, Osio sotto	4,2

Data Collection & Insights

Dataset	Location	Nr Record	Nr Record with Reviews	Completeness	Average Rating Score	Average Compound* (Sentiment)
Conad _superstore_tradotte	Trezzo sull'Adda (BG)	1200	378	32 %	4,2	0,47
Recensioni_ Unes _tradotte	Trezzo sull'Adda (BG)	676	206	30 %	4,1	0,46
Esselunga _tradotte	Stezzano (BG)	1704	909	53 %	4,3	0,42
Recensioni_ Iperal _Presezzo_tradotte	Presezzo (BG)	1200	401	33 %	4,3	0,41
Recensioni_ Penny _tradotte	Brembate (BG)	514	176	34 %	4,0	0,39
Recensioni_ Lidl _Corso Europa_Osiosotto_tradotte	Osio Sotto (BG)	946	313	33 %	4,2	0,29
Iper _Brembate	Brembate (BG)	1986	1986	100 %	3,9	0,27

*: -1 < Compound < +1

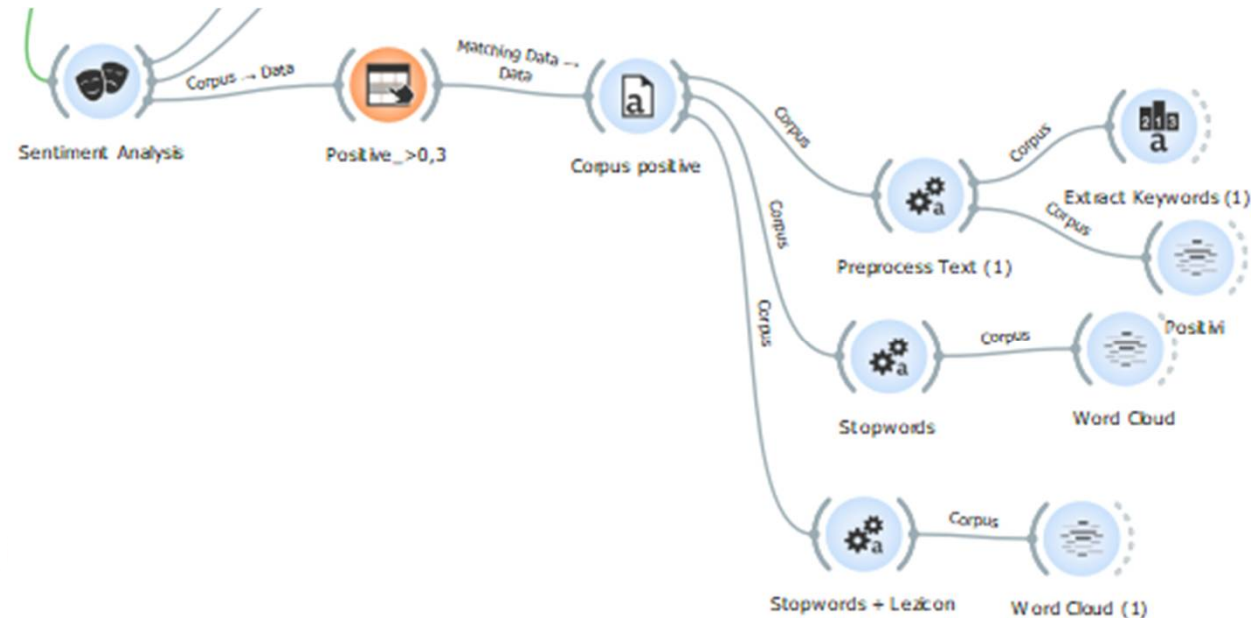
Orange Project Flow - Sentiment Analysis with Vader



Pre-Processing

Applied techniques:

- **Transformation:** url links removal and lowercase
- **Tokenization:** breaking sentences or text into individual words
- **Filtering:** regular expressions & stopwords removal
- **N-grams:** created sequences of n items from a given review to capture context
- **Normalization:** by using UDPipe Lemmatizer, focusing on the meaning of the words



Also tried different pre-processing steps using dedicated **stopwords file** and **lexicon** to better capture key concepts and customer sentiment.

Word cloud Iper Brembate

Negative



Word	TF-IDF
price	0.022
lowest	0.022
lowest price	0.020
area	0.017
price area	0.015
lowest price area	0.014
fuel	0.012
veri	0.012
always	0.012
the	0.012
rude	0.011
staff	0.011
chaotic	0.010
too	0.009
bad	0.009
low	0.009
low price	0.009
servic	0.009
petrol	0.009
dirty	0.009



The negative word cloud shows the sentiment algorithm's limitations regarding semantic and context. E.g.: '*lowest*', '*lowest price*' are interpreted with a negative connotation.

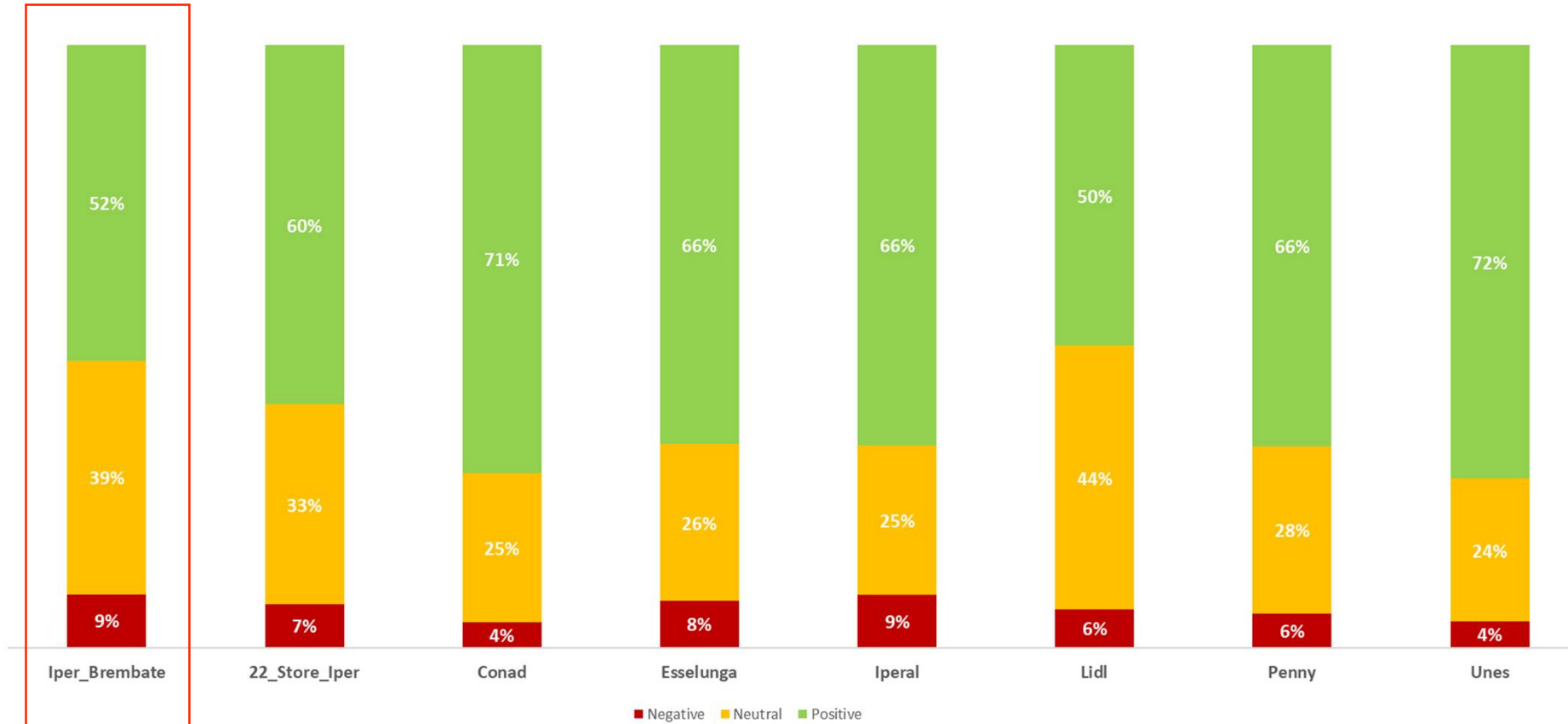
Positive



Word	TF-IDF
good	0.045
optim	0.037
excel	0.030
price	0.025
great	0.025
great price	0.017
excel price	0.013
comfort	0.013
nice	0.012
shop	0.012
veri	0.011
good price	0.011
well	0.011
best	0.011
super	0.011
alway	0.009
everyth	0.009
offer	0.009
stock	0.009

Sentiment Analysis Iper Brembate vs. Competitors

Looking at Sentiment analysis - compared to overall Iper stores – Brembate has less reviews with a positive sentiment (-8%), but also more negative (+2%) and neutral (+6%). Esselunga and Unes (Finiper group) are the most appreciated by customers.



Keywords

Have been identified few main aspects that consumers look at when they shop at Iper Brembate. This list is very similar to the other competitors, and it is worth to deepen the research and run a sentiment analysis for each of the most popular keywords, in order to provide valuable insights for the business.

Keywords were found through TF-IDF function, which measures the importance of a term considering its frequency inside corpus.

$$TF_IDF_{i,j} = tf_{i,j} \cdot \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$: number of occurrences of i in document j

df_i : number of documents containing word i

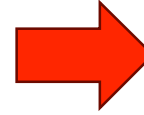
N : total number of documents

Word	TF-IDF
stocked	0.068
quality	0.058
staff	0.037
clean	0.037
parking	0.030
price	0.026
fresh	0.021
friendly	0.018
fish	0.012
quality price	0.011

% Positive Sentiment for Most Important Words Brembate vs. Competitors

Main issues:

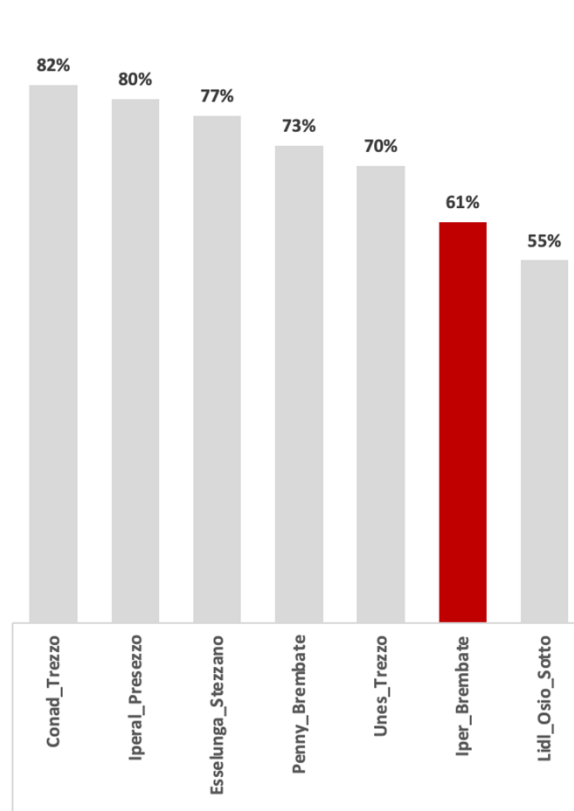
- **Price:** higher than competitors and not always physically visible
- **Products:** often not well stocked
- **Staff:** complaining about kindness



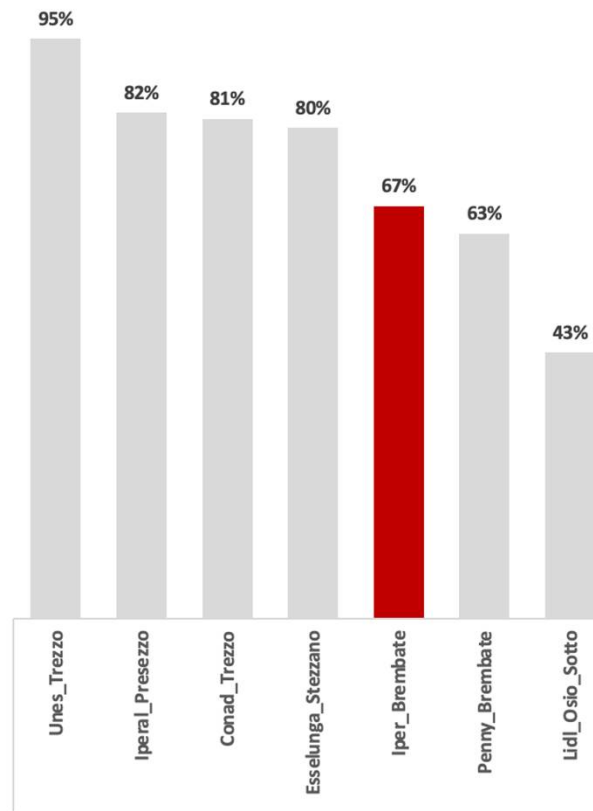
Business insights:

- Work better with Nielsen price data to align price with trading area
- Work with supply chain to analyze out-of-stock reasons
- Create soft skills course with HR

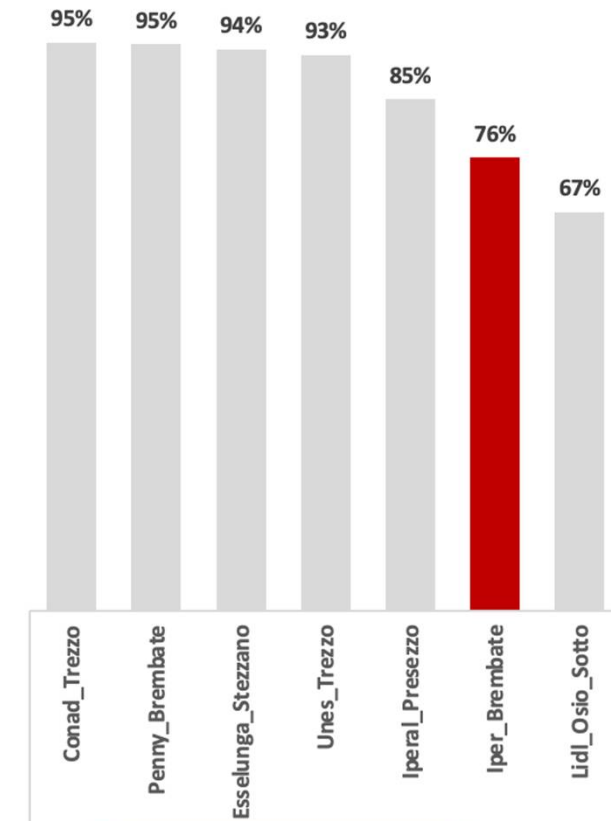
Price



Product



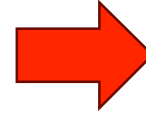
Staff



% Positive Sentiment for Most Important Words Brembate vs. Competitors

Main issues:

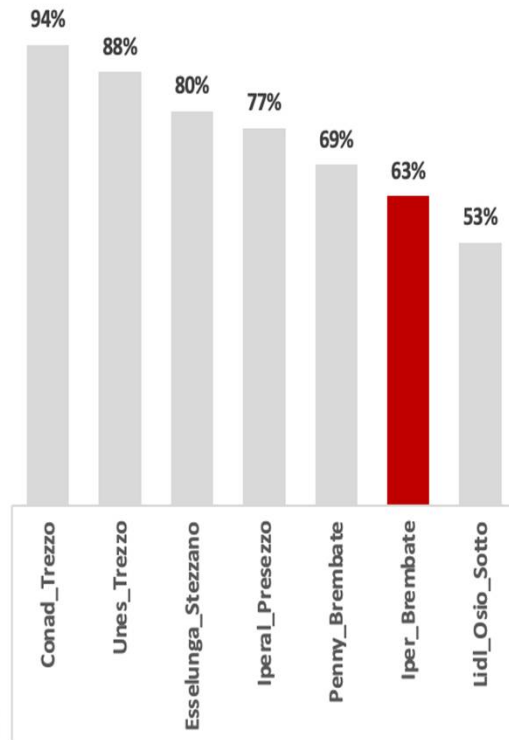
- **Retail space:** chaotic, difficulties to find products
- **Parking:** low availability, issues during the weekend
- **Checkouts:** usually long queues, especially in peak hours



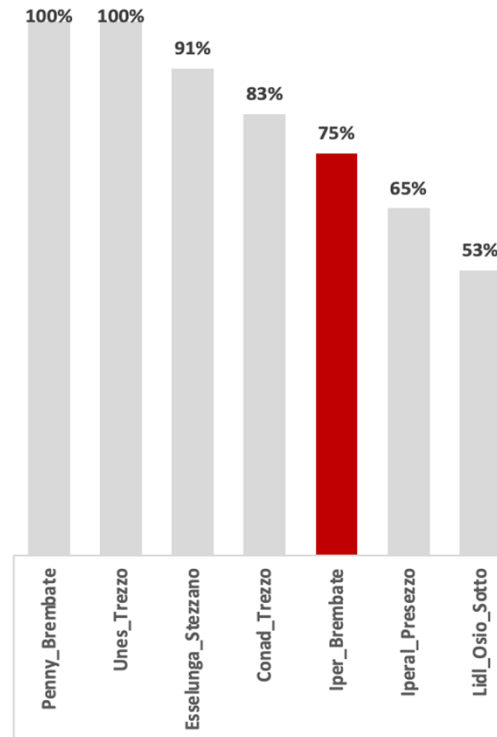
Business insights:

- Create a POC with supplier to analyze layout
- Work with municipality to improve the issue with Leolandia (nearby theme park).
- Create a cross team group to analyze and improve checkouts queue

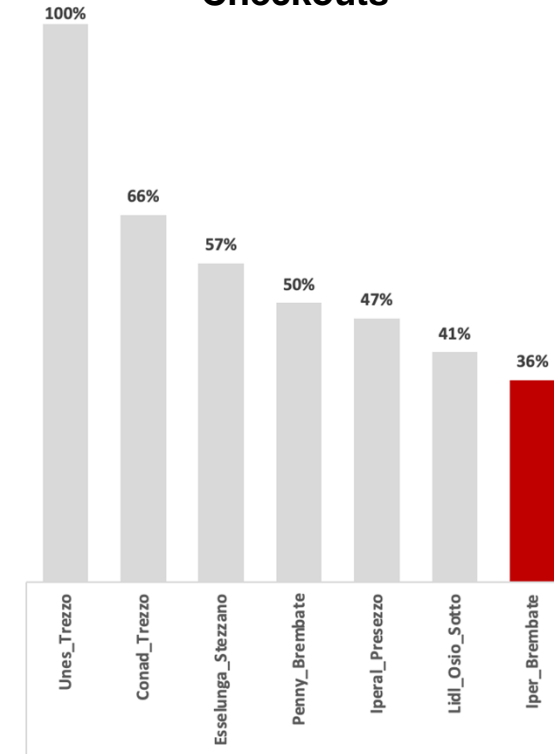
Retail space



Parking

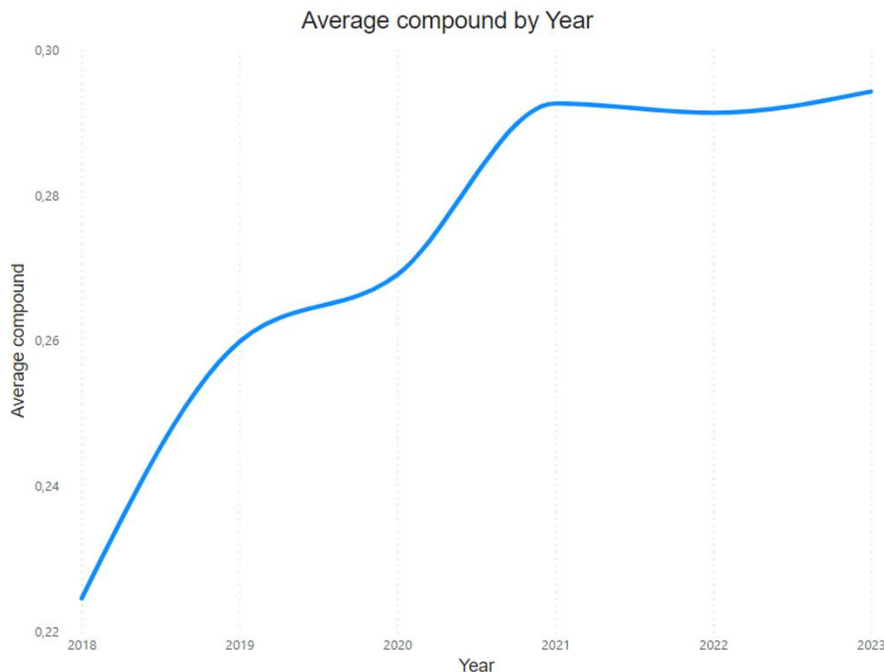


Checkouts

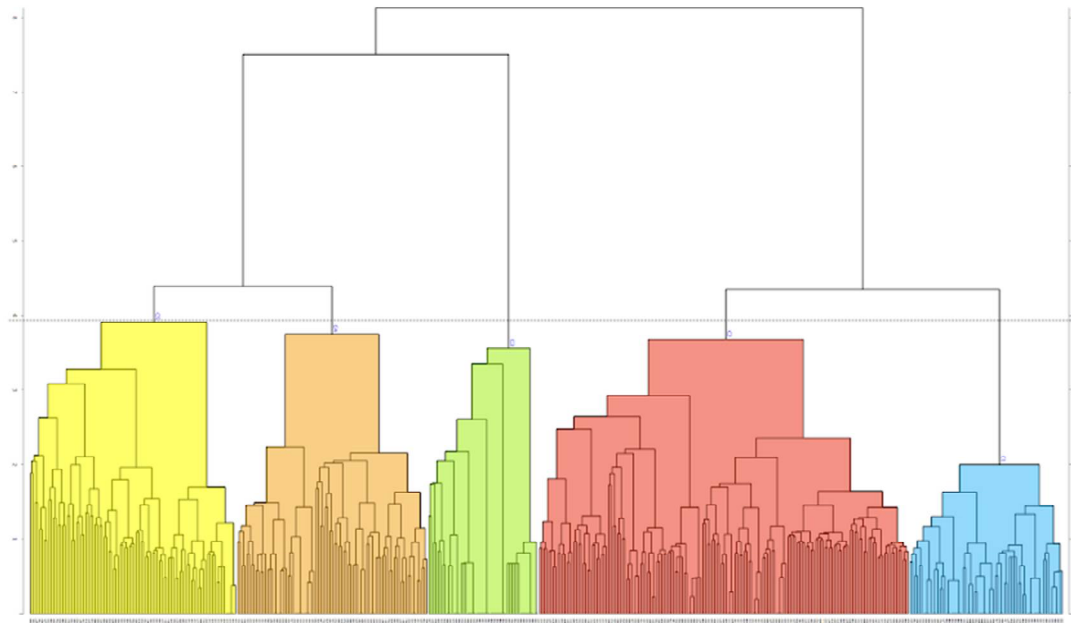
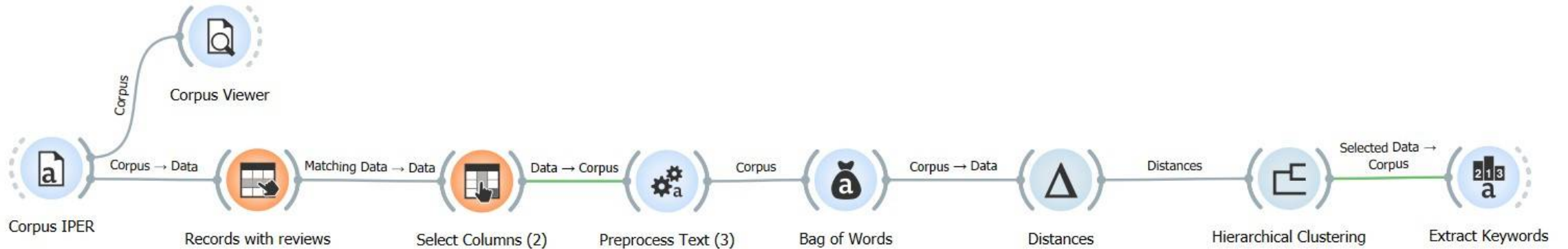


Iper Brembate – Insights over last five years

- Since 2018 the **average compound value** (Sentiment) has increased, but has also reached a **plateau** in the last three years (2021 to 2023). The overall sentiment (0,22 – 0,29) can be considered as '**neutral**'.
- After a steep increase between 2020 and 2021, during 2023 the **average score** is presenting a **strong decline** that needs to be further investigated. (In parallel the number of reviews has decreased too).

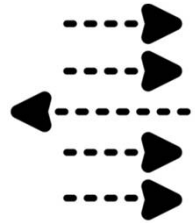


Hierarchical Clustering - Jaccard



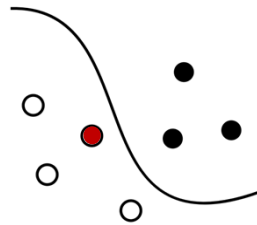
- Trying to create a Hierarchical Clusterization by using Jaccard algorithm, 5 main groups has been returned. (Also tried with different pruning depth but no relevant and reliable results have been reached).
- Unfortunately, by analyzing the qualitative output, this did not satisfy the expectations. System was **not able to find out specific topics** except for adjectives that also seem to be in contrast with each others (e.g. expensive, cheap, optimal, small, ...)

Critical Points



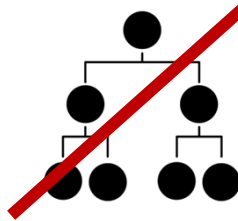
Sentiment-Vader reasoner seems **not always** be able to **distinguish** some words' "weight" **depending on the context**.

E.g.: "*lowest price*" presents a negative compound value! Instead, from the customer point of view, this should be a good thing.



In pre-processing phase even if using an ad hoc stop words list file, the system seems **not recognizing** (so neither removing) **some conjunctions, articles** and **prepositions**.

Only the use of a **lexicon file** was **helpful** to reduce drastically the number of words represented inside the word clouds.



Orange was **not able** to execute a **Hierarchical Clustering** or a Topic Modeling. This is most likely due to the fact the corpus included many comments from a few words to couple of phrases maximum (rather than a verbose document).

Conclusions & Improvement steps



Pain points

- **Retail space:** it is time consuming for customers to easily find all products they are looking for.
- **Products:** promotional products are often out of stock.
- **Price:** higher than competitors and not always physically visible.
- **Checkouts:** poor space management, also highlighted by the long queues at checkouts.



Action points

- Iper, to improve the consumer experience, should conduct analysis with external vendor (e.g. by using AI capabilities to analyze customer journey).
- Company should review the supply chain E2E process to guarantee products availability.
- Analyze competitors' pricing strategy and review the price relevation (by Nielsen).
- Create synergies among teams (IT, Marketing, Sales) as to define checkout lines root causes and possible solutions.

Thank you

Team:

Michele Iannotta | micheleiannotta273@gmail.com

Luca Izzo | l.izzo7@campus.unimib.it

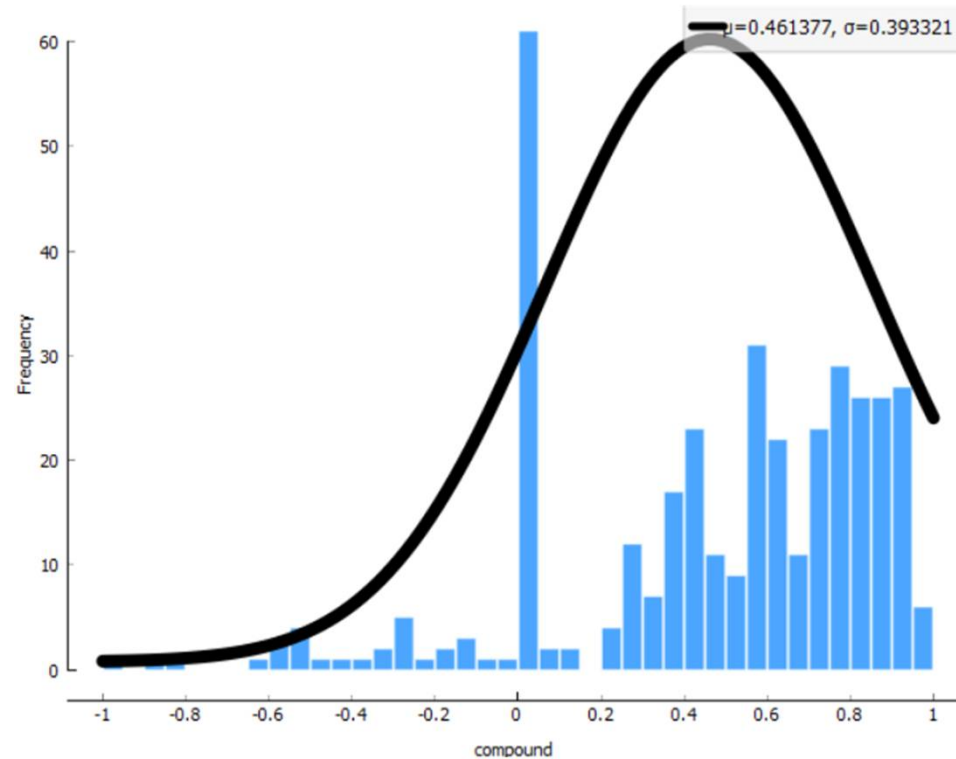
Fabio Jr Lorenzini | fabiojr.lorenzini@gmail.com , <https://github.com/FabioJrLorenzini>

Riccardo Licciardello | riccardonapoli990@gmail.com



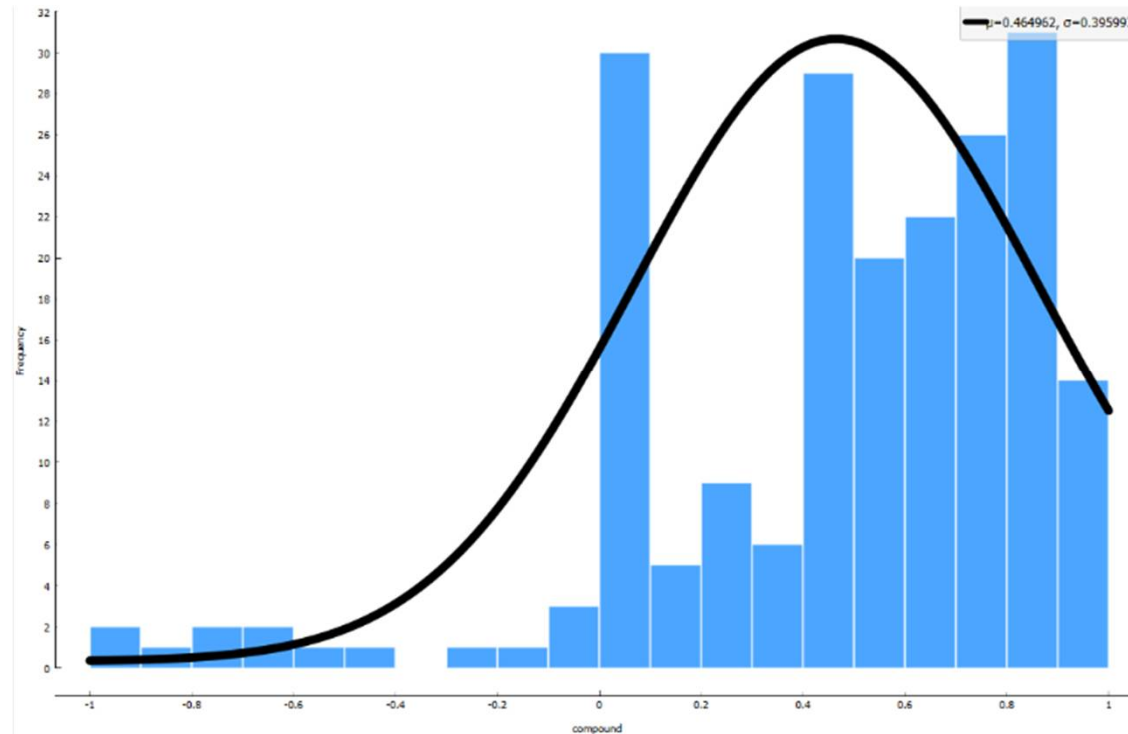
Appendix

Compound distribution – Conad, Trezzo sull'Adda



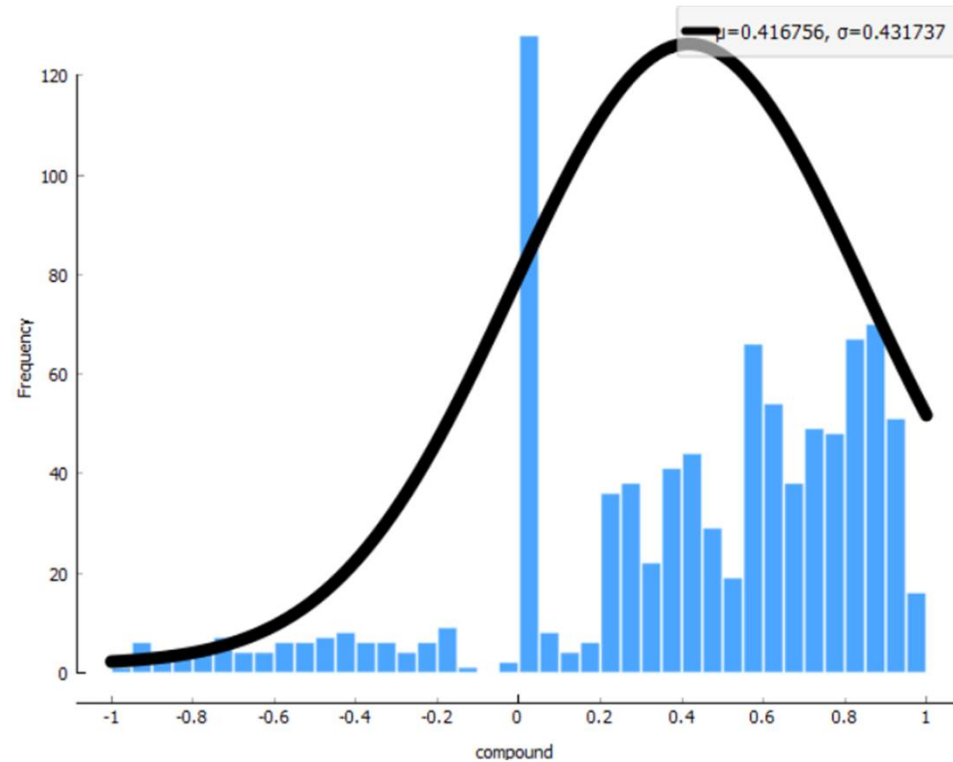
PROS	CONS
Kind staff	Few bad experiences about staff
Clean and tidy spaces	Products promotions not always updated
Good parking availability	

Compound distribution – Unes, Trezzo sull'Adda



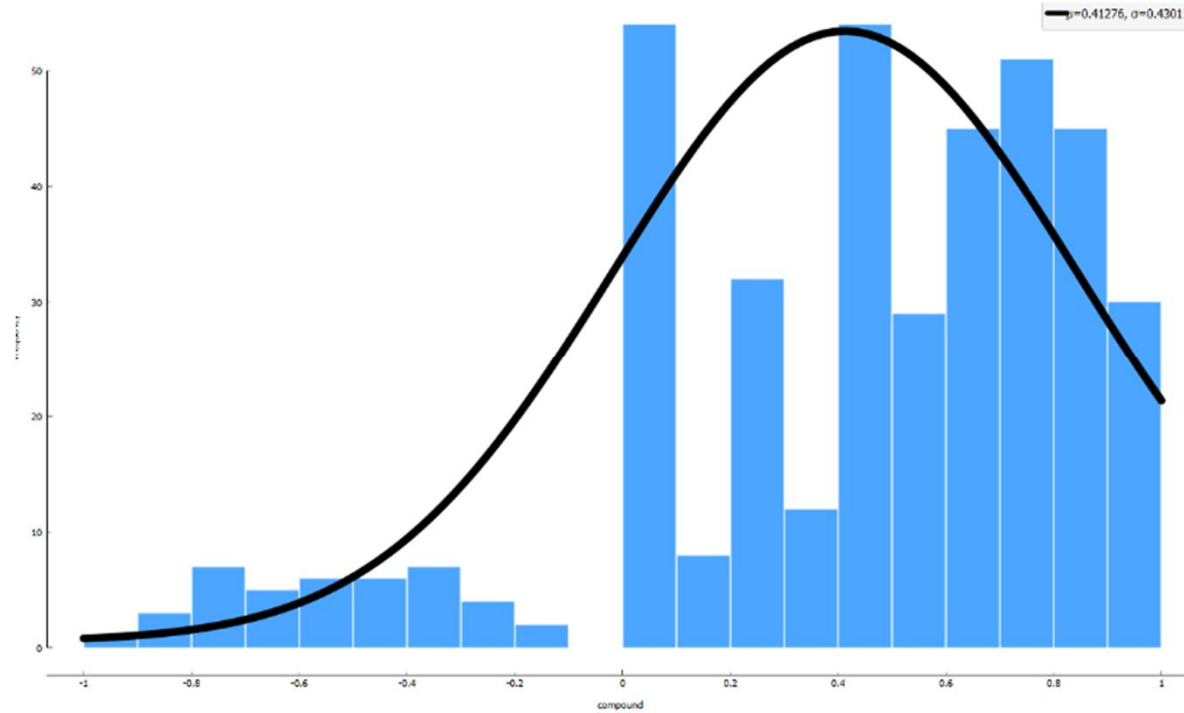
PROS	CONS
Good value for money	Prices seems increasing
Great products choice and quality. Very appreciated 'Viaggiator Goloso' brand	Few reports about products no more available
Very friendly and helpful staff	
10% discount for retired people every wednesday	

Compound distribution – Esselunga, Stezzano



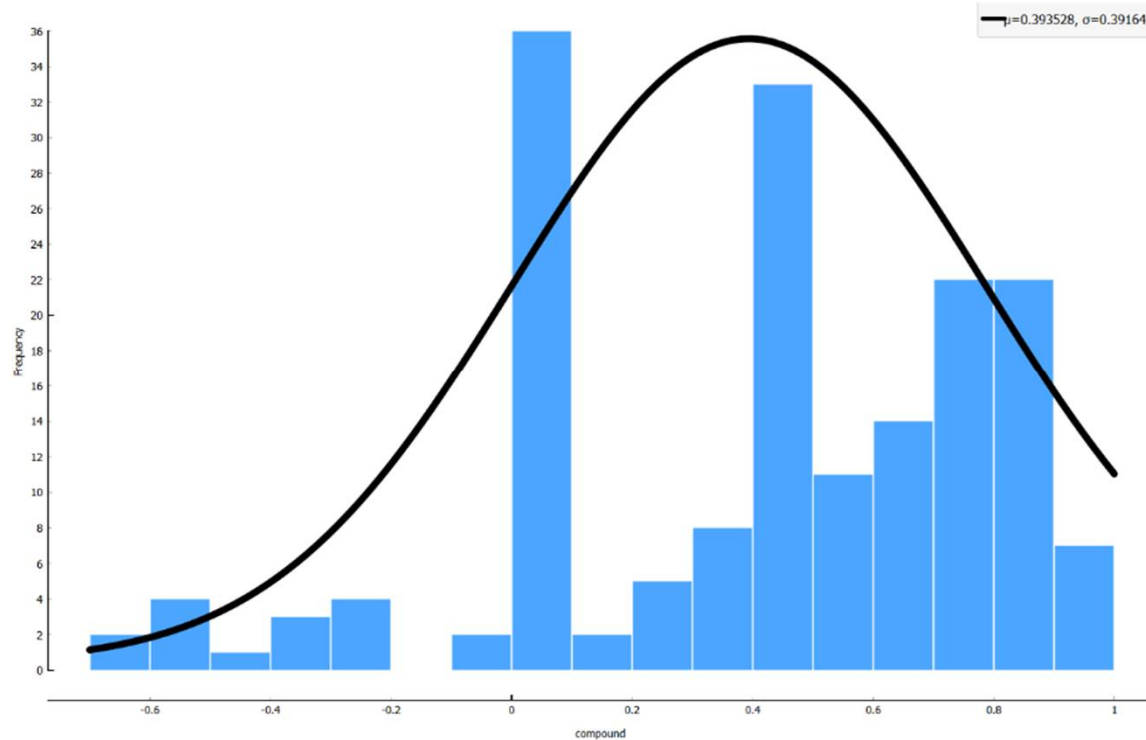
PROS	CONS
High products quality	High prices
Kind staff	
Clean and tidy spaces	

Compound distribution – Iperal, Presezzo



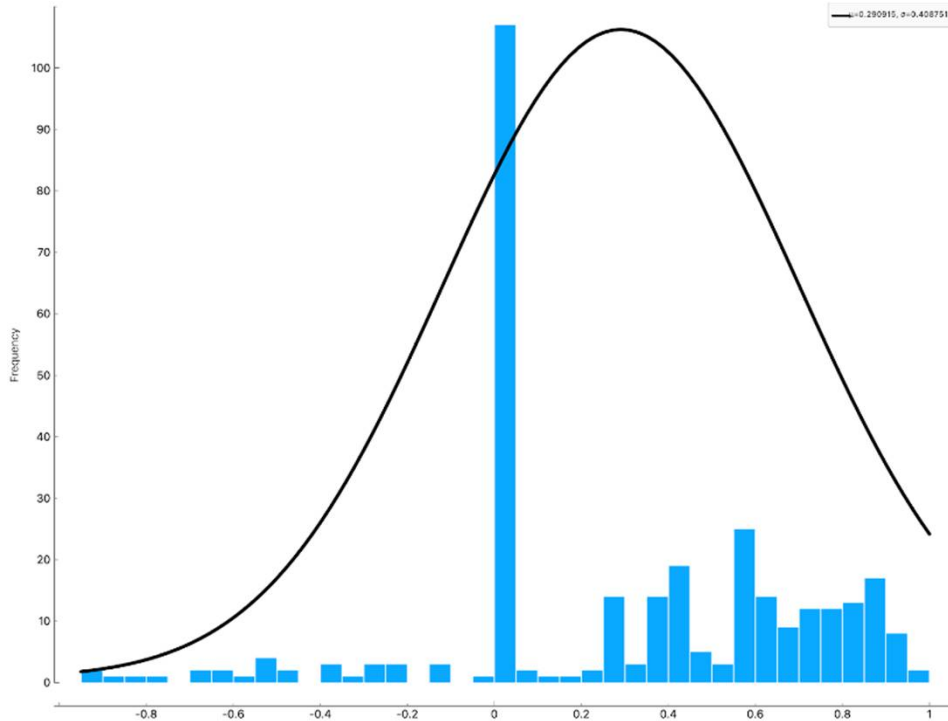
PROS	CONS
Well stocked	Parking is too tight, inconvenient
Very good value for money	Unfriendly staff
High quality and variety of products	Few bad experiences about butcher shop
Tidy and clean spaces	

Compound distribution – Penny, Brembate



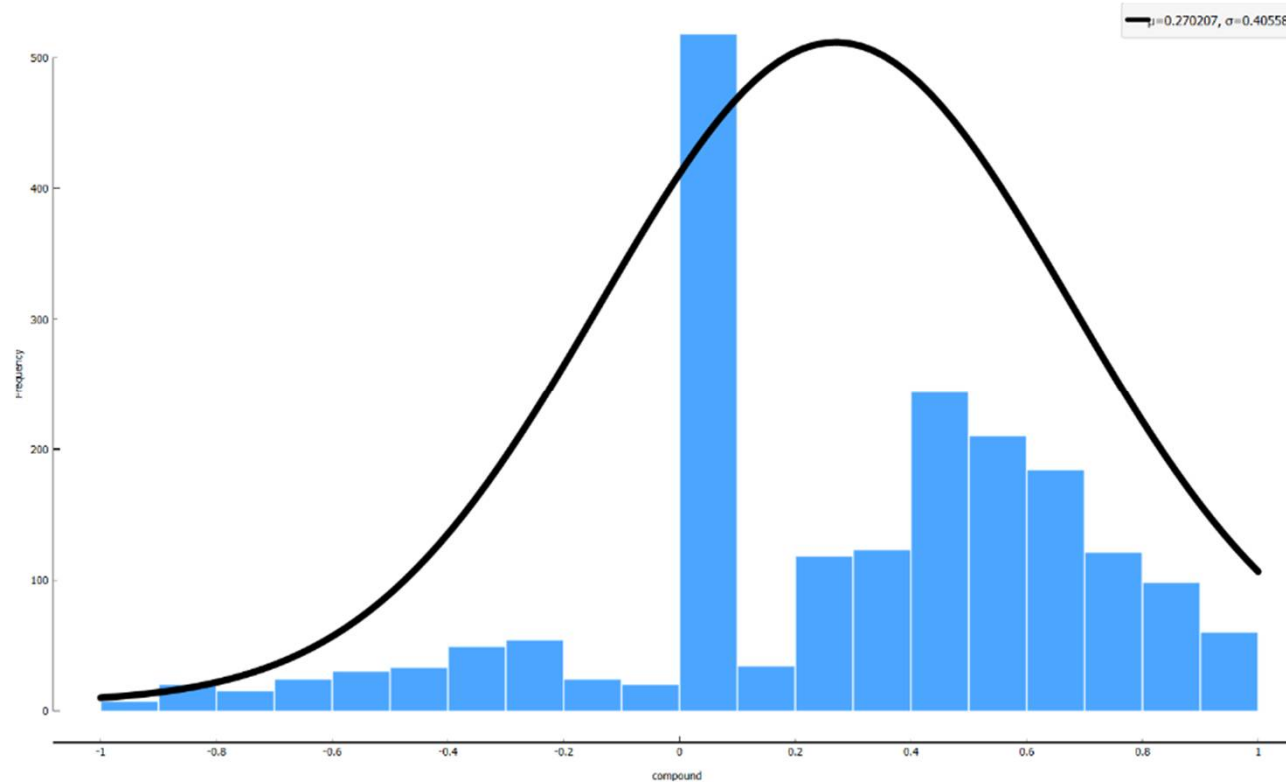
PROS	CONS
Good prices and promotions	Small and messy spaces
Great fruit quality	Lack of products
Very friendly staff	Few complaints about rude cashiers and security
Good parking	

Compound distribution – Lidl, Osio sotto



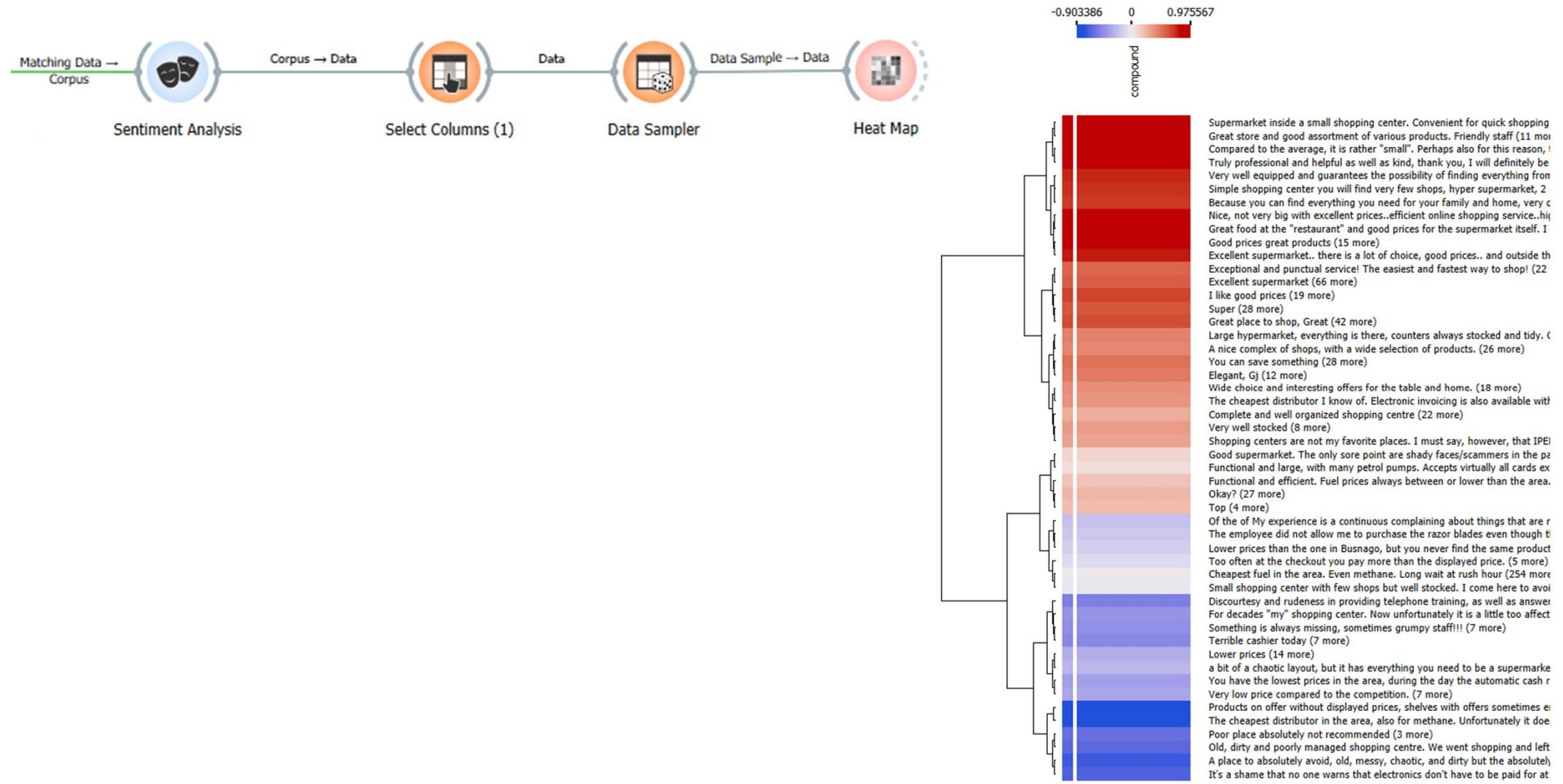
PROS	CONS
Good prices and promotions	Unfriendly staff
Very good quality for vegetables	General bad meat quality
Clean and tidy spaces	Few complaints about very dirty toilets

Compound distribution – Iper, Brembate



PROS	CONS
Good products selection	Prices higher than competitors and price tag not always well visible
Good products promotions	Promotion products often not available
	Poor space management, long queues at checkouts

Sentiment Analysis heatmap – Iper, Brembate



Word cloud Iper Brembate – Pre-processing with Lexicon

Negative



Word	TF-IDF
rude	0.078
staff	0.070
bad	0.067
chaotic	0.053
much	0.053
dirty	0.053
shelves	0.042
messy	0.040
rude staff	0.033
empty	0.032
price	0.032
department	0.025
shelves empty	0.019
discount	0.015
empty shelves	0.013
messy dirty	0.012
department staff	0.009
staff staff	0.009

Positive



Word	TF-IDF
stocked	0.068
quality	0.058
staff	0.037
clean	0.037
parking	0.030
price	0.026
fresh	0.021
friendly	0.018
fish	0.012
quality price	0.011
friendly staff	0.011
cashier	0.010
gastronomy	0.009
meat	0.008
equipped	0.008
staff friendly	0.006
stocked parking	0.004
fresh fish	0.004