

# Sentiment analysis to evaluate the attitudes to target consumers for C.I. HERMECO S.A (OFFCORSS)

Medellín Colombia | TEAM 25

## Business Problem

C.I. HERMECO S.A. (OFFCORSS) is a top tier retailer leading the colombian childrenswear market. This is a company focused on the satisfaction of the customer. For this reason, OFFCORSS has recently created a policy to acquire all data from social networks and customer service in order to understand their customers, inasmuch as the social network has now emerged as an essential part of an individual's life and it has changed the way of live in the 21st century, allowing users share their views about different events, news, and products (Haider et al., 2018). Nevertheless, for this purpose there is no way to trace the information within the acquired data sets. This is the first project of data engineering in the company. Therefore, OFFCORSS requires to establish a robust data process allowing to collect, depurate, process and analyse natural language information from customer support and social networks.

Specifically, OFFCORSS requires a pipeline that acquires, processes and analyses natural language for understanding how their customers feel about their products and competitors. For it, sentiment analysis is an efficient and effective way of finding the people's view, opinion, and the response regarding any product, incident, and an event (Can et al., 2018). Sentiment analysis also helps to computationally find and cluster the views shown in a piece of text (Prabowo and Thelwall, 2009). Taking this into account and given that people post and comment about their views, opinions, and response to different events and products on Facebook regularly, and due to this, Facebook has become a valuable source of sentiments.

Then, OFFCORSS would like to use the result of these analyses to take actions to improve their market, product quality, business processes and customer support.

## Business Impact

OFFCORSS realizes the importance of data to understand their customers. Nevertheless, there are no means for traceability of information of the customer experience with their brand.

Based on the above, we propose the metrics at next to measure the business impact for this project:

### *Business Impact Analysis (BIA)*

KPIs:

- Identify key relationships and dependencies with customer sentiment and the organization profits
- Identify key relationships and dependencies with customer sentiment and business products
- Identify key relationships and dependencies with customer sentiment and business services
- Identify key relationships and dependencies with customer sentiment and the competitors
- Determine possible financial implications of negative incidents
- Determine the type and number of business processes that are threatened by potential negative incidents

### *Risk Analysis and Management*

KPIs:

- Risk assessment process for customer sentiment for each customer comment
- Provide number of risks revealed and mitigated about negative sentiments
- Periodic risk analyses conducted for negative sentiments

### *Incident Response*

KPIs:

- Incident response and management plan
- Link between incident response plan and customer support

- Detailed contact information of critical internal and external contacts
- Process for rapid communication with employees and customers.
- Policies and procedures for dealing with social social media and customers
- Implementation of tracking system

#### *Data Access*

KPIs:

- Regular, frequent backups and scraping of data
- Identify acceptable amount of time needed to get and backup the data

#### *Data Analysis*

KPIs:

- Training of classifiers of text sentiment
- Evaluation of performance for classifier of sentiments

#### *Access to server*

KPIs:

- Determine amount of time required to access files on the server
- Identify acceptable number of errors when processing data

#### *Interface responsiveness*

KPIs:

- Determine amount of time required to visualize different analyses in the interface
- Identify acceptable number of errors when processing the data

## Data

By the company, we have three initial datasets available to perform the analysis:

- **NPS\_Responses:** this dataset contains the responses of a Net Promoter Score (NPS) survey that the company sent to their clients.
  - *Survey Date (Date):* the date when the company sent the survey to the client. In the actual data that we have, there is only the results for a survey sent on 10/08/2020.
  - *Name (Str):* the name of the user assigned to the client responded to. Some of them are the client's name, and others

the name of the process that they have done (like returning a product).

- *User ID (Int)*: an unique identifier of the client that responded to the survey.
- *Email (Str)*: the email address that has sent the response.
- *Rating (Int)*: an identifier in the range 0 to 10 that the likelihood to recommend the company.
- *Classification (Str)*: a categorical variable that clasificate the likelihood to recommend the company into three categories: Promoter (9-10 rating), passive (8-7 rating) and detractor (0-6 rating).
- *Comment (Str)*: the comment that the client made with the survey. Most of the registers don't have a comment associated and some of them have useless values like dots or blank spaces.
- *Response Date (DateTime)*: the date and hour when the client sent the response of the survey.

Though this data contains a Classification column based on Rating, it was found that this might be miss-classified based on customer comments; negative comments classified as promoter, or the opposite. Additionally, there are significant null values in the Comments column. A new classification system based on comments from customers might be more beneficial for the sentiment analysis integrated with data from social networks.

- **Satisfaction\_Ratings**: this dataset contains the level of satisfaction of the client when they have been attended to in any request.
  - *Requester (Str)*: the name of the user who responded to the satisfaction survey. Some of them are the client's name, and others the name of the process that they have done (like returning a product).
  - *User ID (Int)*: an unique identifier of the client that responded to the satisfaction survey.
  - *Email (Str)*: the email address that has sent the response.
  - *Ticket ID (Int)*: an unique identifier of the ticket generated.
  - *Brand (Str)*: the name of the brand associated with the support process. The only brand available in the dataset is "OFFCORSS".
  - *Group (Str)*: a categorical variable that represents the type of the support that the client needs. There are four groups in the dataset: "Call Center", "Soporte OFFCORSS", "Tienda Virtual" and "Venta Directa".

- *Assignee (Str)*: a categorical variable that represents the area of the company that was in charge of response to the support. There are 13 different areas in the dataset.
- *Satisfaction (Str)*: a categorical variable that represents the classification that the client made based on the satisfaction of the solution for the request made. There are only two categories, "good" and "bad".
- *Comment (Str)*: the comment that the client made with the survey. Some of the registers don't have a comment associated and some of them have useless values like dots or blank spaces.
- *Survey Date (DateTime)*: the date and hour when the ticket was generated.

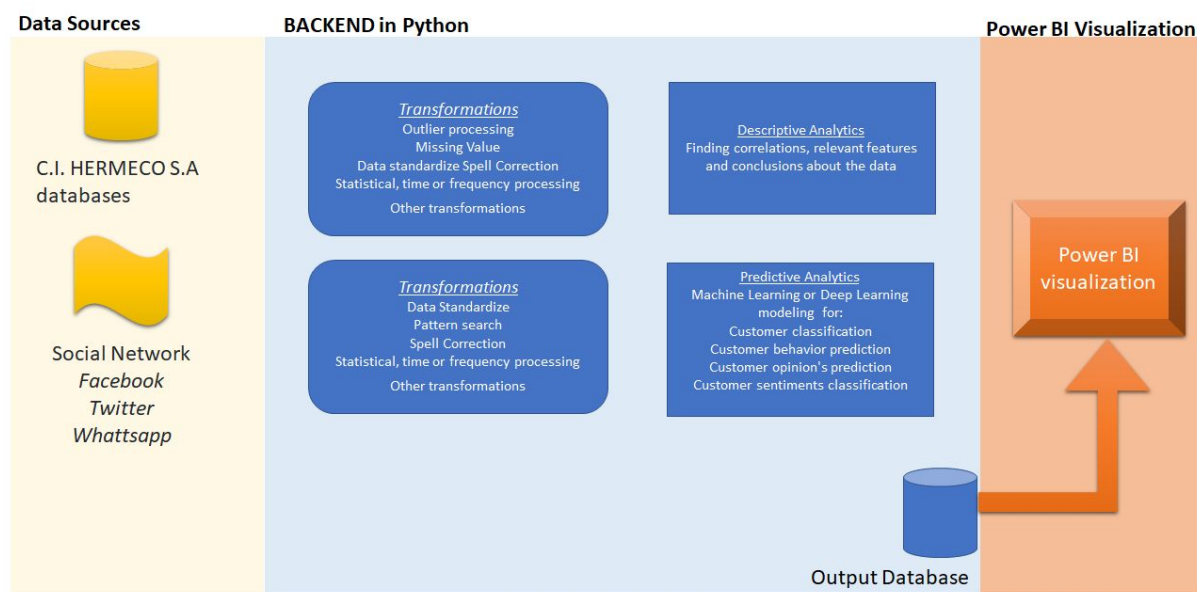
- **OFFCORSS\_Facebook\_Posts:**

- *Post ID (Int)*: the unique identifier of the facebook post.
- *Text (Str)*: the title of the post. It is a character string that can include emojis.
- *Post Text (Str)*: the body text of the post. Like the title, it is a character string that can include emojis.
- *Time (DateTime)*: the date and hour when the post was published. We have available posts from 12/03/2014 to 18/08/2020.
- *Shared Text (Str)*: the text present in the links to other publications that this post shares. Some of them are null when the post doesn't share links to other pages.
- *Image (Str)*: the URL address of the image attached to the post. Some of them are null when the post doesn't include an image.
- *Video (Str)*: the URL address of the video attached to the post. Some of them are null when the post doesn't include a video.
- *Video ID (Int)*: the unique identifier of the video. It is null when the post doesn't have a video associated.
- *Likes (Int)*: the number of like reactions associated to this post.
- *Comments (Int)*: the number of comments that the people have left in the post.
- *Shares (Int)*: the number of the post shares.
- *Post URL (str)*: the Url address of the post.
- *Link (str)*: the URL address of the website that is mentioned in the post.
- *User ID (str)*: an unique identifier of the user account that has published the post.

This dataset represents the most important data for the project. According to OFFCORSS, Facebook interactions have a significant impact on their business and they have realized that the number of interactions in this social network has increased significantly over the past years. Understanding the client perception through the interactions in this social network is crucial for the business health. An advantage from this approach is that the same data obtained from OFFCORSS Facebook page could be obtained for the competitor selected for the analysis. A disadvantage is that it might not contain geographical information.

## Methods

A natural language processing (NLP) system is proposed to characterize C.I. HERMECO S.A. (OFFCORSS) clients' opinions and feelings related to the company. This information is extracted from internal databases and the social networks of C.I. HERMECO S.A. (OFFCORSS) and it's competitors. The general steps in the proposed systems are shown in the following figure.



For data extraction there are two main features:

1. Extracting information from C.I HERMECO S.A (Offcorss): At the time, the company has been able to share csv files with information related to customer surveys and social networks.

2. Web scraping from C.I HERMECO S.A (Offcorss) and it's competitors:  
A web scraping tool will be developed using packages such as Beautiful Soup, Selenium, Lxml and Scrapy from Python. The web scraping process will be developed with the purpose of bringing in the data from the networks, finding patterns in this data and standardizing the data in a readable format.

Once the data is extracted, this will be processed in the backend. The main processing framework will be the following:

1. Processing of missing features: Completing data via interpolation or dropping out rows with missing values.
2. Outlier and noise detections and special treatment for these values.
3. Text processing: Spell correction, Elimination of word connectors, word frequency, adjacent words processing (Collocation), word context (Concordance).
4. Data standardization
5. Tokenization of words or sentences
6. Feature extraction

From the processing outputs it is expected to identify insights, correlations, and conclusions from the data which will be the core of the descriptive analytics. This may include but is not restricted to: Customer clustering, correlation between words and customer perceptions, Principal component analysis in the main features extracted and customers, regional conclusions about the perception of C.I HERMECO S.A (Offcorss) and it's competitors. The previous descriptive analysis are related to the following business questions:

1. What is the general perception of the customers of C.I HERMECO S.A (Offcorss) and it's competitors services and products?
2. Which departments or features can be improved from C.I HERMECO S.A (Offcorss) to increase client satisfaction?
3. In which regions should C.I HERMECO S.A (Offcorss) focus its efforts to increase client visibility, satisfaction and profits?

4. In which periods of time should C.I HERMECO S.A (Offcorss) increase advertising to increase profits?
5. Which products produce the greatest impact in the clients and in the revenues of the company?.

The data can be split into type of clients, date and regions to answer the previous questions. The most relevant graphs that we are going to use are scatter plots to identify correlations, outliers and groups in the data. Bar plots and line plots (for timeseries) to identify patterns in time and heat maps to identify the impact in the different regions. Finally, for the processing will be used tools like natural language toolkit or the algorithms will be implemented by the team group in Python.

For the predictive stage, we will build machine learning or deep learning models such as support vector machines, random forest, deep neural networks with convolutional or LSTM (long short term memory) layers for the classification and regression tasks. The final selection of the predictive models will depend on the quantity and quality of the data. If required, we will also consider implementing optimization algorithms such as gradient descent, Newton's method or genetic algorithms. This will depend on the final availability of the data and the performance of detection algorithms.

## Interface

The processing results will be stored in a relational database and this database will be connected to a Power BI report. The most relevant interaction between users and data will be filtering and displaying results..

In the Power BI interface we want to display the following information:

1. C.I HERMECO S.A (Offcorss) and it's competitors client classification and perception by region
2. C.I HERMECO S.A (Offcorss) and it's competitors influenced by date and products.
3. Relationships between services and client satisfaction.
4. Sentiment prediction and classification due to C.I HERMECO S.A (Offcorss) actions.
5. Most common words in comments and posts by clients categories.
6. If possible, most relevant products from C.I HERMECO S.A (Offcorss) and impact on their clients.



# Milestones

The project milestones have been organized according to the consecution of the main objectives. It is important to mention at this point, that given how generic each of the work units can be, the group decided to define the scope of work for each milestone, containing the bare minimum technical requirements. Additional features may be added or will just be collected as a list of suggestions for further development, according to the availability of resources.

## **1. Data Scraping**

Conclusion of the first milestone should result in the construction of a dataset containing the collection of the messages, mentions and/or interactions in general available for the brand and its clients. Data scraping will be focused on the acquisition of information published in Facebook. Currently, Facebook is listed among the most widely used social networks. People post and comment about their views, opinions, and response to different events and products on Facebook regularly. Due to this, Facebook has become a valuable source of sentiments.

If possible, data scraping efforts would be extended not only to OFFCORSS, but also to its main competitors in Colombia, as this could lead to identifying key pieces of information to develop strategic actions.

## **2. Text analysis**

Nurtured by the result of the previous milestone, a series of routines will be performed in order to organize, clean, standardize and enable the analysis performed line by line, looking to extract key information relevant to the brand positioning on the market place.

## **3. Sentiment analysis**

Having processed the dataset compiled and prepared, sentiment analysis workflows will be performed in order to “understand” the general consent of the brand, its products, services, stores and in general, interpret the general opinion of the brand according to the data available through social media.

Results of basic NLP workflows applied to the organized dataset captured from multiple sources of data.

#### **4. Tool wrap-up and publication**

Having accomplished all the workflow required for the data acquisition, processing and analysis, the team will proceed to organize all the code containing the logics to perform the before mentioned steps in a way that facilitates the knowledge transfer to the OFFCORSS staff.

#### **5. Competitors analysis**

The main intention of this milestone is to take advantage of the tools developed by the team 25, specially designed to analyze the public perception of the brand OFFCORSS, in order to extend its capabilities, to also capture the general impression of the public to its identified most close competitor in the Colombian territory.

This will provide key insights of the second company to OFFCORSS, that most likely will use such information to put together a corporate strategy to distance themselves from their competitors and expand the company's footprint in the nationwide marketplace.

## **Timeline**

Weeks 1-2:

- Project description. Done!

Week 3-4:

- Project scope. Done!

Week 5:

- Dataset source: Web scraping Facebook OFFCORSS.
- Dataset source: Web scraping Facebook Competitor.
- Dataset source: NPS Responses (include more surveys).
- Dataset source: Satisfaction Ratings (include more surveys).

Week 6:

- Basic EDA: For all datasets. Prioritize which ones will be used and perform data cleaning on those.

- Define if additional datasets are required for the project. If so, request those to OFFCORSS.

Week 7:

- In-depth EDA: For all datasets.
- Jupyter notebook with EDA.
- Front end design. Power BI definition.

Week 8:

- Front end design completed. Power BI.
- Back end infrastructure design.

Week 9:

- Front end infrastructure completed.
- Databases hosted in the cloud.

Week 10:

- Application completed.

## Concerns

- Geographic information available? Location of stores and clients with bad perception?
- Interactions in Instagram social networks are important for business analysis and client understanding. Web scraping from Instagram to be defined.


Zoom Meeting

You are viewing Maria Lara Cuervo's screen

View Options

Enter Full Screen


# Concerns



What are the primary concerns you have at this point?

- Data that is not available / hard to get
- Gaps of information between your team and the organization
- Data analysis/modeling skills

Anything that you think can become a potential issue for the success of the project!



Unmute

Start Video

Participants 25

Chat

Share Screen

Record

Leave


Zoom Meeting

You are viewing Maria Lara Cuervo's screen

View Options

Enter Full Screen

# Final remarks



- Focus on big-picture strategic issues as much as on day-to-day activities
- Avoid duplication of effort and make sure everyone is clear about who is doing what
- Seek and give each other constructive feedback
- Maintain a *can-do* approach when you encounter frustrating situations
- Appreciate one other's unique capabilities

