LUISS 'Guido Carli'
MSc in Data Science and Management - Data Science in Action Course
May 5, 2024

The dark side of "Chiara" Ferragni

Pandoro-Gate: an online focused analysis

Project Technical Report

Marcoccia Andrea, Team Leader
ID: 775451, andrea.marcoccia:@studenti.luiss.it

Costantini Giorgia
ID: 782261, email: giorgia.costantini@studenti.luiss.it

Navarra Filippo
ID: 782571, email: filippo.navarra@studenti.luiss.it

Collaborative Business Case Proposed by Deloitte

1 Introduction

The scandal known as the "Pandoro Gate," involving influencer Chiara Ferragni and confectionery company Balocco, has captured public attention and sparked widespread interest. Rooted in the collaboration between Balocco and Chiara Ferragni for the production of the "Pink Christmas" pandoro, apparently to support children hospitalized at Regina Margherita Hospital in Turin, the initiative quickly came under fire with hefty multimillion-dollar fine for unfair commercial practices followed.

The Antitrust Authority sanctioned Chiara Ferragni and Balocco for their 2022 Christmas promotional campaign, because Ferragni's companies, Fenice and TBS Crew, along with Balocco, advertised the "Pink Christmas" pandoro (priced at two and a half times the cost of non-branded pandoro), implying to consumers that a portion of the purchase would be donated to the Turin hospital for osteosarcoma and Ewing sarcoma research and to purchase new equipment. However, Balocco had already made the donation to Regina Margherita Hospital in May 2022, well before the Christmas season. Therefore, Ferragni's companies, despite collecting over €1 million through the initiative, allegedly failed to contribute (Ferraiuolo, 2023).

After two days of silence, the influencer publicly apologized, first through an Instagram story and later with a video posted among her social media content. But how did the media react to this scandal? What were the consequences for Chiara Ferragni?

In this report, we will analyze how the public opinion reacted to the incident using Natural Language Processing (NLP) techniques. We will examine the impact on the Ferragni brand and propose potential strategies for recovery.

2 Methods

2.1 Instagram

Data Collection
The data set for this analysis consists of comments extracted from eight Instagram posts made by Chiara Ferragni around the time of the 'Pandoro Gate' scandal. These posts were strategically selected to capture the sentiment evolution of Ferragni's fans before, during, and after the event. The selection criteria were twofold:

- **Temporal Relevance:** Posts were chosen based on their dates to create a timeline that encapsulates key moments of the scandal. This temporal distribution ensures a chronological understanding of sentiment shifts.
- **Content Consistency:** To minimize the content-induced sentiment variability, posts with similar themes and subjects were selected. This homogeneity allows for a more accurate assessment of sentiment changes driven by external events rather than variations in post content.

The comments were collected using the web tool 'exportcomments.com', which facilitates the extraction of up to 100 comments per post in CSV format, making the data gathering process both efficient and systematic. The specific dates of the posts analyzed are: 27-09-2023, 25-10-2023, 18-11-2023, 14-12-2023, 18-12-2023, 04-03-2024, 17-03-2024, and 23-03-2024.

Data Pre-processing
The initial datasets consisted of 100 comments per post. The key pre-processing step involved:

- **Cleaning of Nested Comments:** Comments containing '@' symbols, which typically indicate responses to other comments rather than to the post itself, were removed. This was essential to ensure that the analysis focused solely on direct reactions to the posts.

This cleaning process resulted in a reduction in the number of usable comments per post, impacting the dataset size as follows:

Post's Date	From	To
27-09-2023	100	83
25-10-2023	100	80
18-11-2023	100	62
14-12-2023	100	56
18-12-2023	100	49
04-03-2024	100	63
17-03-2024	100	81
23-03-2024	100	69

Table 1: Reduction in the Number of Comments

This refined approach ensures that the sentiment analysis reflects only the direct engagements of users with the content, providing a more accurate representation of the public sentiment towards Chiara Ferragni during the specified timeline.

2.2 Facebook

Data Collection
For the analysis of public sentiment on Facebook regarding a specific case, comments from posts made by three prominent newspapers—La Repubblica, Corriere della Sera, and Il Giornale—were selected. These posts were identified as among the most commented on this topic on Facebook, making them particularly suitable for a comprehensive sentiment analysis. As of the date of this report, April 24, 2024, the posts had accumulated a significant number of comments: 1590, 1222, and 610, respectively and comprehensive of nested comments. To efficiently gather these comments, we utilized a web scraping tool from apify.com, specifically the 'Facebook Comments Scraper'. This tool is designed for robust data extraction from Facebook, allowing developers to automate the collection of public reactions on posts. For the purpose of our project, the tool was configured to extract only direct comments from the selected posts, ensuring the relevance and consistency of the data obtained.
Data Pre-Processing

The initial extraction yielded a total of 365, 269, and 181 comments from the respective posts. To prepare the data for analysis, our primary pre-processing step involved handling missing values, which were present only in the dataset extracted from the third post (Il Giornale). Null values were identified and removed, resulting in a cleaned dataset of 170 comments for this post. The data from the other two posts did not require additional cleaning beyond the removal of null values, as the configuration of the scraping tool effectively filtered out irrelevant or indirect comments. This ensured that our analysis would be based on direct user interactions, providing a clear and focused insight into public sentiment.

2.3 Reddit

Data Collection and Pre-Processing
To analyze Reddit user sentiment regarding the 'Pandoro Gate', we selected three highly commented posts on Reddit, drawing from discussions hosted in the subreddits r/Italia and r/Italy. As of April 24, 2024, these posts had accumulated a total of 411, 324, and 274 comments, respectively, inclusive of nested comments. These posts were chosen for their relevance and the volume of engagement, providing a rich dataset for sentiment analysis.

The links to these selected posts are as follows:

- Sfogo su Chiara Ferragni
- Chiara Ferragni e il caso Balocco - Chiedo scusa e...
- Antitrust maxi multa a Chiara Ferragni e Balocco

For data extraction, we utilized PRAW (Python Reddit API Wrapper), a Python package facilitating access to Reddit's API, enabling the automated collection of comments.
Using PRAW, we implemented a specific script configuration to interact with the Reddit API. Here's a breakdown of the process:

1. **API Authentication**: Configured the Reddit object with necessary credentials to authenticate our access.
2. **Post Selection**: Utilized the unique post ID to retrieve each targeted discussion.
3. **Comment Extraction**: Configured the script to replace nested comments with a limit of 0, which ensured the extraction of only top-level comments directly attached to the main post.
4. **Data Structuring**: Iterated over each top-level comment to construct a dataset containing the original post text and the comment text.

This method enabled us to extract 144, 98, and 62 top-level comments from each post, respectively. By focusing only on direct responses to the original posts, we ensured that the sentiment analysis would be directly relevant to the primary content discussed, rather than secondary discussions occurring in nested replies.
No further pre-processing was needed.

2.4 News Articles

Data Collection
News articles were sourced from Google News using the *requests-html* library, which facilitated the web scraping process. We conducted searches on Google News with specific keywords to target articles relevant to our study. The process was detailed as follows:

1. **Session Initialization and URL Definition:** We initiated a session using the AsyncHTMLSession from the *requests-html* library and defined our URLs based on the following keyword searches:
   - Chiara Ferragni Balocco
   - Ferragni Balocco (yielding some additional different results)
   - Pandoro Gate
2. **Rendering the Page:** We used the .arender() method to render the page in a headless browser. Parameters were set to scrolldown=5 to load more articles and sleep=1 to ensure all dynamic content was fully loaded, thereby preventing errors and ensuring a comprehensive scrape.
3. **Article Selection:** We targeted the general article HTML tag to locate the news content on the rendered page.
4. **Data Extraction:** The specific data we were interested in (article titles and links) were located within anchor tags with the class a.JtKRv.
5. **Data Storage:** Using a for loop, we iterated through each article, extracting and storing the title and link in a list for subsequent analysis.

This systematic approach allowed us to efficiently collect and organize data directly related to our research keywords, providing a structured dataset for further analysis.

The next step involved getting the actual text from these links. For this task we decided to use the *newspaper3k* library automates the extraction and parsing of news articles from the web. It downloads articles, extracts main text and other contents[1]. It supports multiple languages, allowing for good results also with non-English websites.
We extracted the text article from the links previously found. The output was stored as; Title, Text of the Article and URL.

This approach combines the strengths of both libraries, web scraping with requests-html to handle dynamic content retrieval and deep content extraction and processing with newspaper3k.

[1]It also removes the clutter (such as navigation bars, ads, footers) that often comes with web pages.

2.5 Sentiment Analysis through Mistral AI Apis

API Configuration
We used two distinct models from Mistral AI depending on the complexity of the analysis required:

1. **Open-Mistral-8x7b**: This is an open source LLM, that uses 12.9B active parameters out of 45B total. The costs of using this model are very low, but is still very powerful for many tasks.
    - Input Cost: $0.7/1M tokens
    - Output Cost: $0.7/1M tokens
2. **Mistral-Small**: This is a private optimized model and is ideal for cost-efficient, low-latency operations, ideal for straightforward tasks like classification of text.
    - Input Cost: $2/1M tokens
    - Output Cost: $6/1M tokens

Each model was carefully selected to match the complexity and specific requirements of the text data being analyzed, ensuring an efficient allocation of computational resources and minimizing costs.

API Usage
Our interaction with the API began by setting up a client object with our API key, enabling secure communication with Mistral's servers. The main task involved creating prompts tailored to guide the model in classifying sentiments.

Sentiment Classification Workflow:

1. **System Message Setup:** We defined a system message that served as an initial guide for the API, outlining the task of sentiment classification.
2. **Prompt Preparation:** For each comment, we formulated a unique prompt that directed the model to classify sentiment. The label requested from the model varied depending on the social media platform being analyzed, reflecting the nuanced differences in language use across platforms.
3. **API Interaction:** With the prompts prepared, we invoked the API, sending batches of messages and collecting the generated responses. Each response was then parsed to extract the classified sentiment.
4. **Error Handling:** Any API errors encountered during interactions were logged and handled gracefully to maintain data integrity and consistency across the dataset.
5. **Result Compilation and Storage:** Extracted sentiments were compiled and appended to their respective datasets. The final results were saved in a CSV format to facilitate easy access and reuse, avoiding the need for repetitive and costly API calls.

This methodical approach to using Mistral AI's APIs allowed us to process and classify sentiments from multiple sources efficiently. The next sections will detail the specific configurations and refinements made to the prompts, outputs, and model selections for Instagram, Reddit, and Facebook, highlighting how tailored approaches were implemented to suit each platform's unique data characteristics.

2.5.1 Instagram Configuration

For Instagram, where our objective was to assess sentiment regarding Chiara Ferragni under her posts and observe changes over time, we opted for a straightforward classification approach. The API was tasked with determining if each comment was 'Positive', 'Neutral', or 'Negative'.

API Setup:

- **System Message**: Established to guide the model: Classify the sentiment of the following Instagram comments as 'Positive', 'Neutral', or 'Negative'.
- **Prompt Formatting**: Each comment was presented as follows: Classify the content of this Instagram comment, replying only with 'Positive', 'Neutral', or 'Negative': comment
- **Model Selection**: We utilized the open-mixtral-8x7b model due to its efficiency and cost-effectiveness, which was adequate for the relatively simple task of classifying Instagram comments.

The Facebook comments under articles about the 'Pandoro Gate' scandal required a nuanced approach to accurately capture diverse sentiments. The comments' classification included labels such as 'Contro Chiara Ferragni', 'Felice della notizia', 'Neutrale', and 'A favore di Chiara Ferragni'.

API Setup:

- **System Message:** Crafted to contextualize the analysis based on the article's content, guiding the model to classify comments into one of four categories. The message included a brief description about the article and instructions on how to categorize different sentiments:

  > L'articolo è il seguente: {titolo dell'articolo}. Classifica il sentimento dei seguenti commenti all'articolo su Facebook in una delle 4 categorie: *'Felice della notizia'*, *'Neutrale'*, *'A favore di Chiara Ferragni'* o *'Contro Chiara Ferragni'*. Un commento che esprime gioia o felicità va classificato come 'Felice della notizia'. Un commento neutrale va classificato come 'Neutrale'. Un commento che esprime supporto a Chiara Ferragni va classificato come 'A favore di Chiara Ferragni'. Un commento che critica Chiara Ferragni va classificato come 'Contro Chiara Ferragni'.

- **Prompt Formatting:** Adjusted to align with the defined categories, ensuring clarity in the classification task:

  > Classifica il sentimento di questo commento, rispondendo solo con 'Felice della notizia', 'Neutrale', 'A favore di Chiara Ferragni' o 'Contro Chiara Ferragni': {comment}.

- **Model Selection:** Given the complexity of discerning nuanced sentiments from text, we employed the *Mistral-Small* model, which is better suited for detailed reasoning and handling context-heavy tasks.

2.5.2 Reddit Configuration

The configuration for Reddit is equal to the Facebook one, using the same nuanced categorization to capture sentiments in comments discussing Chiara Ferragni's controversy. The only 2 differences are that the context is the one of the title of reddit posts, and the comments given in input are the ones of reddit posts.

These configurations permitted us to achieve great results, with sentiment classification accuracy for the topic being superior to that of other sentiment analysis methods like pre-trained models.
Additionally, executing the file 'sentiment_benchmark.ipynb' Mistral-Small model's accuracy can be verified for an external dataset of labeled sentiment data on Twitter comments, where the model achieved an accuracy of around 56%. Given the context provided in our API calls, we are confident that we achieved superior accuracy in sentiment classification for our dataset.

Moreover, utilizing a Large Language Model (LLM) for this task offers advantages beyond accuracy. It enables us to handle different languages and emojis effectively. Without an LLM, comments containing emojis might have been removed and comments in non-English languages translated, risking the loss or alteration of valuable information in the dataset. By leveraging an LLM, we can maintain the integrity of the data while accurately capturing sentiment across various languages and emotive expressions.

2.6 Latent Dirichlet Allocation for News

LDA is a statistical approach for discovering themes, or "topics," within a large collection of text documents. It does so by identifying patterns in the occurrence and co-occurrence of words. It starts by randomly assigning each word to a topic and an iterative refinement process where each word's topic assignment is continuously optimized, based on; the prevalence of the topic in the document at hand and the word's affiliation with the topic across the full document corpus.

We based our analysis on the coherence score, which evaluated the quality and intepretability of our generated topics. The choice of parameters has also considered the narration of topics.
For our specific news article case about Chiara Ferragni - Balocco, we used the *pyLDAvis.gensim models* library. An initial coherence score has been computed and gave us the results contained in Table 2, which led to the choice of number of topics.

Table 2: Coherence Values for Different Numbers of Topics

Num Topics	Coherence Value
2	0.3684
3	0.5773
4	0.5326
5	0.5603
6	0.5323
7	0.5008
8	0.4905
9	0.5016
10	0.4959

We decided to work on $n = 3$ topics and 20 passes[2]. The number of passes has been set manually checking the outcome. We then considered words than appeared at least in 10 articles and not above the 50% of all articles. This filtering step is a way to remove words that are too rare to be meaningful or too common to be distinctive.
The λ parameter is used to balance the importance given to a term's frequency within a topic against its exclusivity to that topic when ranking terms. We then decided to set $\lambda = 0.75$, which is a common choice that enhances the readability of topic content.

3 Results and discussion

Our sentiment analysis of comments from Instagram shed light on the public reaction to the Ferragni-Balocco scandal(Figure 1). Prior to the scandal, sentiment towards Chiara Ferragni appeared predominantly positive, but a notable shift occurred with a peak in negative sentiment observed on the last post before the scandal erupted. This negative sentiment persisted and even intensified following Ferragni's "Apology Video" three days after the scandal broke. Interestingly, comments were restricted until the Fazio Interview, after which the level of negative sentiment decreased to nearly previous levels, maybe for the loss of interest in the case.

[2]the model will pass over the entire corpus 20 times during training

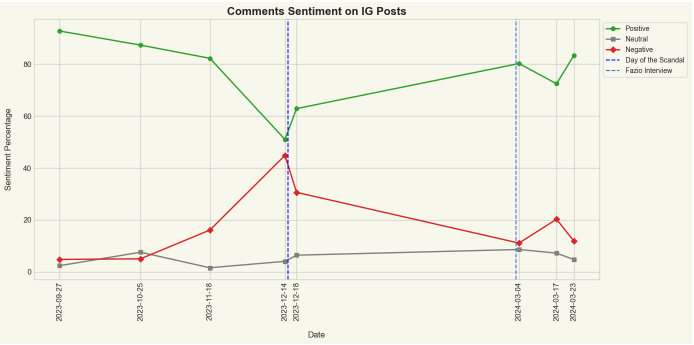


Figure 1: Instagram Sentiment

As for both Facebook (Figure 2) and Reddit(Figure 3), the majority of comments were classified as neutral, probably because of the long comments and more descriptive ones, with a minimal percentage showing support for Ferragni and a higher percentage expressing negativity towards her, often celebrating her "conviction".

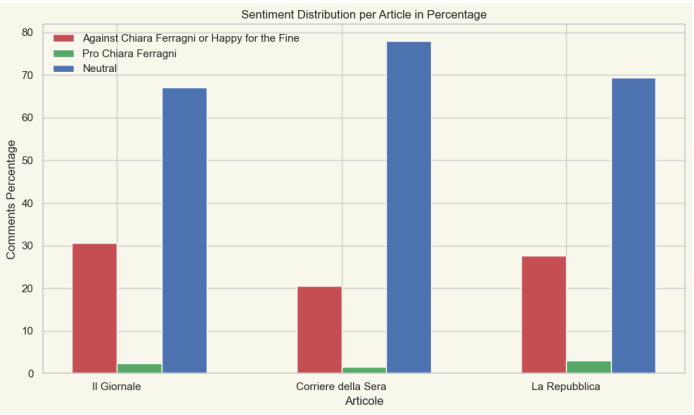


Figure 2: Facebook Sentiment

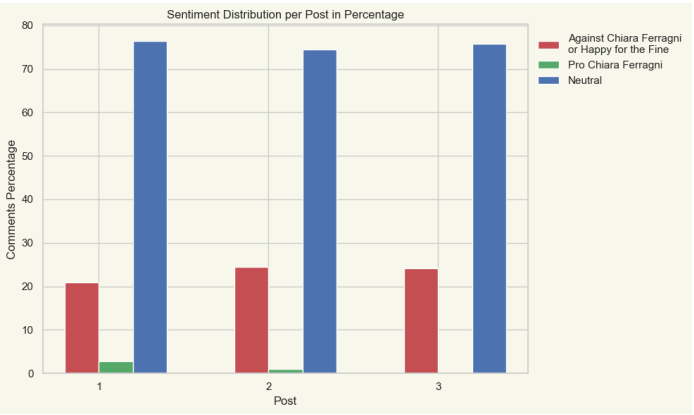


Figure 3: Reddit Sentiment

Furthermore, our topic modeling analysis of newspaper articles underscored the multifaceted nature of the scandal's discourse, revealing three principal topics:

- **Topic 1 - Commercial Theme**; is the largest one, contains 39.4% of tokens. The topic includes terms like "sales", "branded" and "Antitrust" and "Fenice Srl"[1], which suggest the focus on the commercial activities of Miss Ferragni.

Figure 4: Word Cloud for Commercial Theme

- **Topic 2 - Social Media and Public Life Theme**; with 33.8% of tokens. The presence of terms such as "Fedez", "video", "follower", "Instagram" and "communication", indicates a topic centered around social media and therefore her public life.

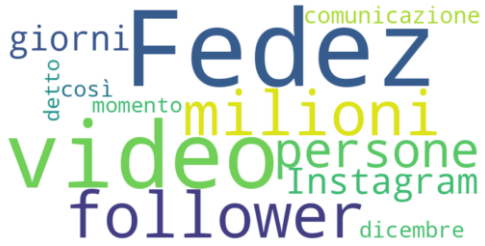


Figure 5: Word Cloud for Social Media and Public Life Theme

- **Topic 3 - Legal Theme**; is the smallest topic with 26.9% of tokens. Terms like "fraud", "Prosecutor's Office", the name of the prosecutor "Eugenio Fusco", "crime", "Codacons" and "suspects", suggest the related legal matters the case involved.

Figure 6: Word Cloud for Legal Theme

These themes illustrate the numerous linkages between the dispute and various socioeconomic spheres and capture the range of considerations surrounding it.
This model, with the specific parameters adjusted for our case (i.e. number of passes, number of top terms), outcome a coherence score equals to 0.6513, improving initial values.

[1] licensor of Chiara Ferragni brands

3.1 Economic impact: Ferragni's losses

The scandal surrounding Chiara Ferragni has lead to a crisis within her business empire, marked by allegations and disruptions in key partnerships, including with Stafilo, Coca-Cola, and Monnalisa (Bianchi, GiovaniReporter, February 2024). In response, the influencer has opted for a strategy of silence. However, the sustainability of this approach is now under exploration, as highlighted in a report by the Arcadia agency authored by Domenico Giordano (Arcadia Mood, March 2024). Giordano's analysis of Ferragni's social media presence revealed a notable decrease in her follower count, totaling 515 thousand individuals out of a pool of 29 million. While seemingly negligible, this decline assumes significance when considering the economic value attributed to each follower even though Giordano notes the absence of data regarding the engagement levels of these lost followers, casting uncertainty on their impact. Furthermore, the substantial reduction in content output has disrupted follower habits, consequently diminishing the account's success, which prompts the algorithm to penalize the reach of the content.

Giordano asserts that Ferragni may soon need to reconsider her silence strategy if she wishes to sustain her brand's presence in the content economy. The diminishing engagement on her Instagram profile further underscores the urgency of this reassessment, with engagement dropping from 5.8% in January 2023 to 0.54% in January 2024, declining further to 0.18% in March (Il Tempo.it, March 2024) (Figure 7). Moreover, the financial implications of her silence strategy are noteworthy, as each sponsored post potentially yields €93,000 in value (Giordano, Marzo 2024). Based on this, it is possible to estimate that by reducing her monthly sponsored post count to one in March 2024, from an average of 23 in September 2023 (before the scandal), Ferragni's potential earnings dropped from over €2 million to around 90K (Figure 8).

Given these circumstances, there arises a pressing need for Ferragni to undertake a brand identity reconstruction, shifting towards values not affected by scandal, such as her family relationships, which seems to be what she is doing at the moment. Additionally, using scandals that potentially attract more attention, such as her divorce from her husband Fedez, could serve as a strategic distraction from the primary controversy (Bianchi, GiovaniReporter, February 2024).

2

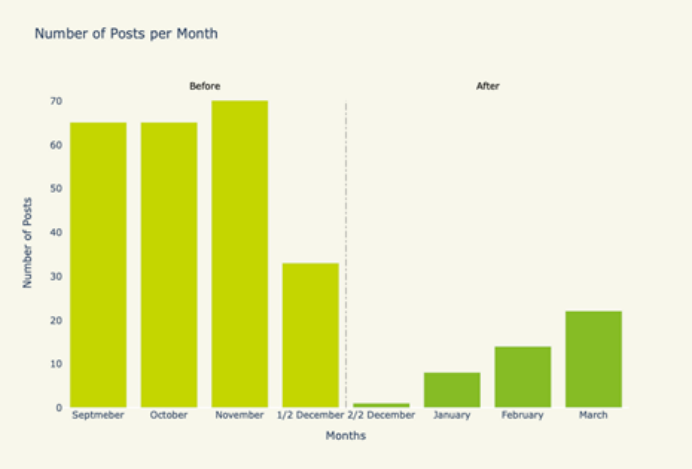


Figure 7: Monthly number of Instagram posts before and after the scandal

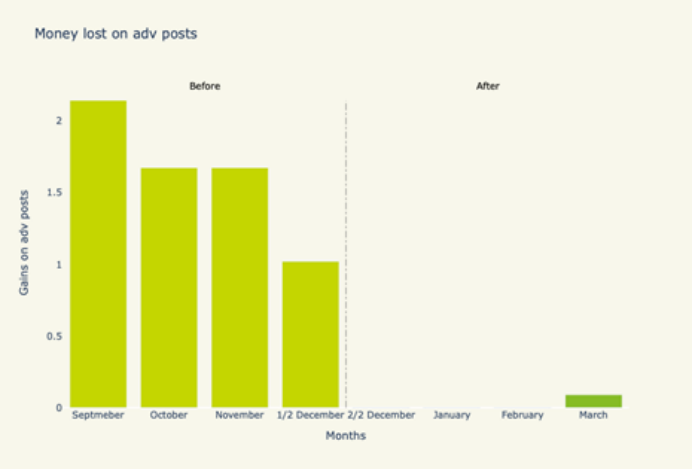


Figure 8: Gains in millions for sponsored posts before and after the scandal

4 Conclusions

Our examination of the Ferragni-Balocco controversy highlights the challenges associated with maintaining a public image in the era of social media, shedding light on how public opinion might change in response to a scandal, which can result in a decline in popularity and an increase in hostility toward the central character. In order to successfully handle these obstacles, Ferragni needs to put an emphasis on openness and cautious brand management. She also needs to interact with her audience in a genuine way to win back their trust and modify her brand identity to satisfy changing customer needs.

By putting these strategies into action, Ferragni can start to lessen the effects of the incident and work toward earning back her previous reputation and public image.

A Appendix

In the Python code, we start by installing necessary packages and importing essential libraries for data analysis and visualization.

Instagram Data
Instagram comments related to Chiara Ferragni posts are extracted through exportcomments.com, imported from CSV files, cleaned by nested comments, and concatenated into a single DataFrame. Sentiment analysis is conducted using the Mistral model`open-mixtral-8X7b`, categorizing comments into either positive, negative, or neutral sentiments. The sentiment distribution over time is visualized using a line plot.

Facebook Data Collection

Moving to Facebook, comments under three different journal articles are scraped using apify.com, loaded from CSV files, cleaned, and sentiment analysis is performed using the Mistral model `mistral-small-latest`. Comments are categorized into predefined sentiment categories: Against Chiara - Pro Chiara - Happy for Fine - Neutral. The sentiment distribution is visualized then for each article with a bar chart.

Reddit Data Collection and Analysis

Similar steps are repeated for Reddit posts. The PRAW library is used to access Reddit's API and scarpe Data of comments under 3 most commented posts on Reddit on the matter. Mistral APIs are then called with the same procedure of Facebook classifying comments either as Against Chiara, Pro Chiara, Happy for Fine or Neutral.

News Articles Data Collection and Analysis

The process starts by gathering news articles from Google News using `requests_html`. By navigating through the HTML structure, the section containing URLs and titles of news articles is identified and collected into a list for further processing. Then, newspaper3k extracts text from these URLs. Next, Latent Dirichlet Allocation (LDA) is applied using `gensim.models` to uncover topics in the articles. Finally, `pyLDAvis` visualizes the LDA results, aiding in exploring the topics.

Chiara Ferragni's Economy Data Analysis

Finally, data on Chiara Ferragni's economy are manually collected and organized into a DataFrame. Plots are generated using Plotly to visualize the decrease in published posts and the financial loss before and after the scandal.

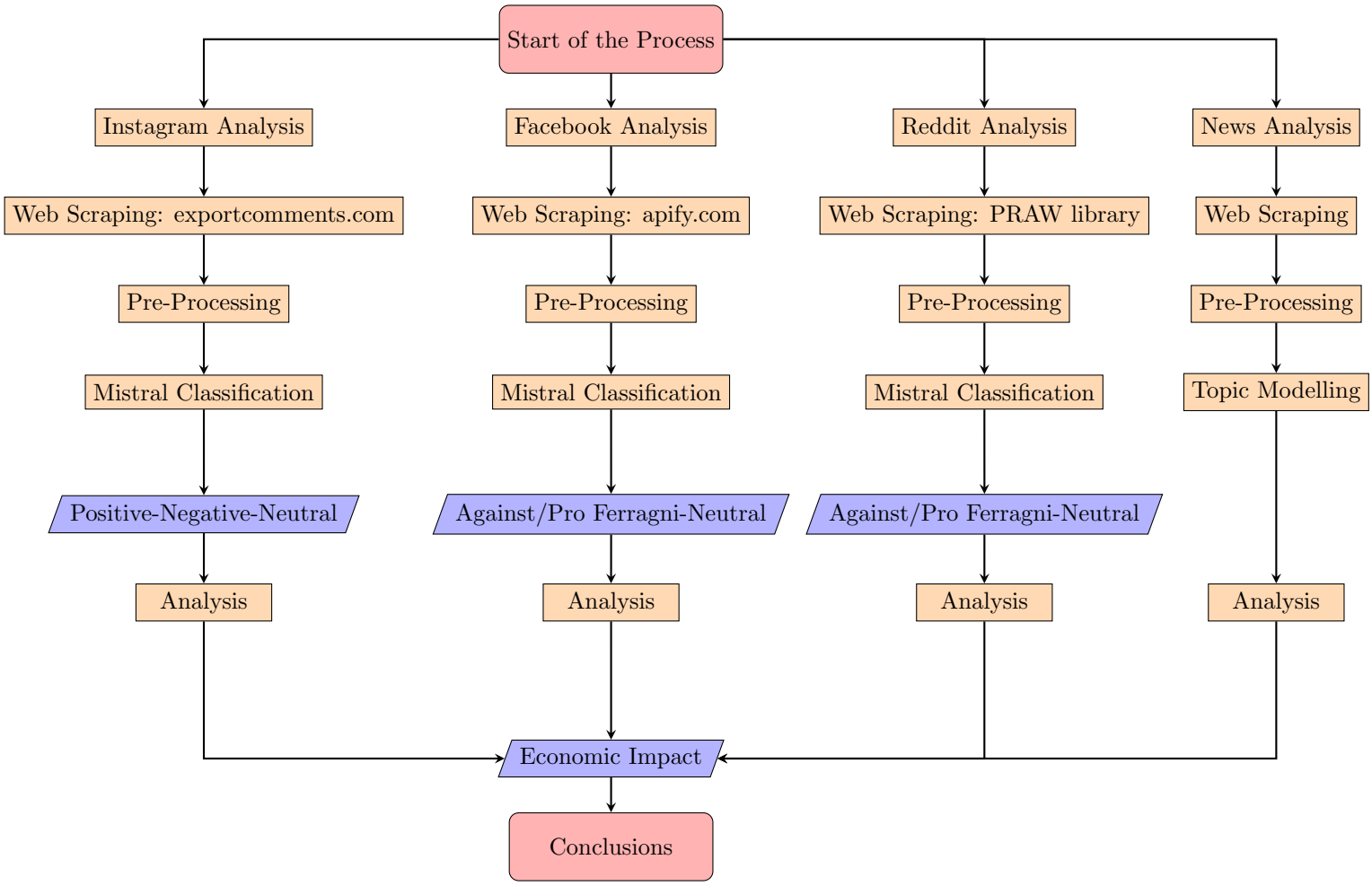
B Appendix

Credit author statement

Giorgia Costantini: Conceptualization, Methodology, Software (Reddit), Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing, Visualization

Andrea Marcoccia: Conceptualization, Methodology, Software (Instagram and Facebook), Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing, Visualization, Supervision

Filippo Navarra: Conceptualization, Methodology, Software (News), Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing, Visualization.


Start of the Process
Instagram Analysis
Facebook Analysis
Reddit Analysis
News Analysis
Web Scraping: exportcomments.com
Web Scraping: apify.com
Web Scraping: PRAW library
Web Scraping
Pre-Processing
Pre-Processing
Pre-Processing
Pre-Processing
Mistral Classification
Mistral Classification
Mistral Classification
Topic Modelling
Positive-Negative-Neutral
Against/Pro Ferragni-Neutral
Against/Pro Ferragni-Neutral
Analysis
Analysis
Analysis
Analysis
Economic Impact
Conclusions