Pandoro Gate: Shedding Light on Sentiments

Public Sentiment and Brand Analysis

Brauner J., Grande P., Luzi S.

Index

1. Introduction......3

2. Methods......3

   2.1. Data Collection......3
   2.2. Data pre-processing......4
   2.3. Models......5

      2.3.1. Keyword Extraction......5
      2.3.2. Topic Modelling......6
      2.3.3. Entity Recognition......7
      2.3.4. Sentiment analysis......8

3. Results......9

4. Discussion and Conclusions......10

1. Introduction

This report delves into the media frenzy ignited in September 2023, surrounding the Italian influencer Chiara Ferragni and the esteemed Italian confectionery company, Balocco, famously dubbed the 'Pandoro-Gate.' The crux of the matter lies in the 'Pink Christmas' affair, wherein Chiara Ferragni and Balocco find themselves embroiled in an investigation for suspected aggravated fraud. This centers around the promotion of branded pandoro, purportedly associated with charitable donations to the Regina Margherita hospital. The Antitrust Authority has intervened, levying penalties against the implicated parties for engaging in unfair commercial practices. Their scrutiny has underscored a glaring incongruity between the advertised philanthropic endeavors and the actual magnitude of contributions. Consequently, a collective fine totaling 1.495 million euros has been imposed on the implicated entities." In this scenario, "the legion of idiots", as Italian writer Umberto Eco referred to those who have gained the right to speak through social media, has provided ample data to conduct a comprehensive analysis of public opinion regarding the aforementioned case.
Our goal is to leverage advanced NLP techniques to understand not only the prevailing sentiment but also the emerging thematic trends shaping public discourse. Our approach combines data collection methods, such as web scraping, with cutting-edge models for text processing and analysis, including sentiment analysis, entity and keyword extraction, and topic modeling. This report elucidates the methods used for data collection and preprocessing, describes the NLP analysis techniques implemented, and discusses the obtained results, highlighting both technical insights and business-oriented implications, particularly focusing on how our findings may have practical and strategic implications for businesses and for understanding broader cultural and social phenomena.

2. Methods

In this Methods section, we delve into our approach for this study, covering data collection, pre-processing, and model implementation. We detail the process of gathering data, explaining why specific datasets were chosen. Additionally, we outline the steps involved in pre-processing raw data; each decision is justified by its impact on data quality and model performance. Furthermore, we discuss the models employed for analysis, such as keyword extraction, entity recognition, topic modeling, and sentiment analysis, providing insight into the rationale behind their selection.

2.1. Data Collection

As mentioned earlier, our data collection primarily revolved around harnessing the vast reservoir of information available on social media platforms. We meticulously extracted data from a variety of sources, including Instagram, Facebook, Quora, Twitter, YouTube, and Reddit. This multi-faceted approach allowed us to capture a diverse range of perspectives and insights from different corners of the digital landscape.

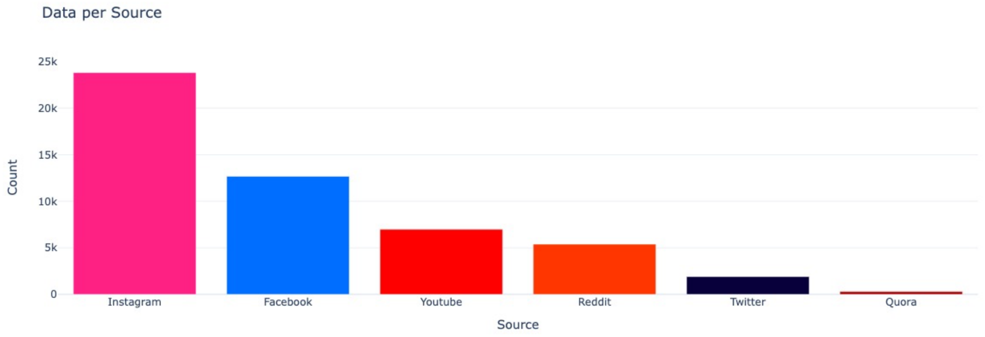

Figure 1, "Data per Source"

Figure 1, a visual representation of our data collection efforts, offers a glimpse into the sheer magnitude of information amassed from each platform. Notably, Instagram and Facebook emerged as dominant contributors, collectively constituting a significant portion of our dataset. Instagram and Facebook alone yielded a combined total of 36,495 entries, representing approximately 71% of the entire dataset. Furthermore, substantial contributions were observed from YouTube, Reddit, Twitter, and Quora, further enriching our dataset with diverse perspectives. The remaining portion of the dataset is distributed as follows: 6,979 entries from YouTube, 5,387 entries from Reddit, 1,904 entries from Twitter, and 292 entries from Quora. Each of these platforms contributes valuable insights, adding depth and diversity to our dataset. In essence, the raw dataframe comprises a total of 51,057 entries.

While determining our data collection methodology, we carefully considered the structural nuances of each platform. We opted for DataMiner in contexts where posts, questions, and comments were prevalent, as its functionality aligns well with extracting data from such formats. Conversely, BrowserFlow proved more suitable for platforms characterized by individual posts. Our approach was strategic, tailored to maximize the efficacy of each bot based on the platform's layout and content organization. Consequently, DataMiner was deployed for Instagram, Facebook, and Quora, while BrowserFlow was utilized for Twitter, YouTube, and Reddit. This tailored approach ensured efficient data extraction across diverse social media platforms, optimizing our dataset for comprehensive analysis.

2.2. Data pre-processing

The very first step of our preprocessing involved merging the various CSV files obtained from each of the sources mentioned in the previous section into a single dataframe. Subsequently, we conducted an exploratory data analysis (EDA) on this merged dataframe to gain insights and inform further preprocessing steps. Our initial file csv consists of 8 columns:

1. Source: Indicates the platform from which the data originated (e.g., Instagram, Facebook, Quora, Twitter, YouTube, and Reddit).
2. Username: Represents the commenter's username.
3. Title: Provides the title of the content (e.g., video title for YouTube, post content for Facebook, or specific tweet for Twitter).
4. Text: Contains the actual text of the comment.
5. Date: Denotes the date when the comment was posted.
6. Url: Provides the URL link associated with the content, facilitating reference to the title.
7. Likes: Indicates the number of likes received for each comment.
8. Post_ID: Primarily utilized for Reddit, serves as a unique identifier for the post.

Initially, our dataset comprised 51,057 comments. As first steps in the preprocessing phase, we opted to remove null text entries, thereby reducing the dataset size to 50,719. Subsequently, we further refined the dataset by removing duplicate entries, resulting in a final count of 49,635 comments. Later, we developed a custom function to identify and extract emojis from the text data and, to enhance the comprehensiveness of our preprocessing, we integrated two dictionaries (abbreviations_italian and emoji_italian) into our script. This allowed us to replace symbols and abbreviations not recognized by the basic function. We then standardized text data by replacing emojis and abbreviations with their corresponding textual representations, we removed URLs, HTML tags and multiple spaces, ensuring uniform text formatting. The resulting Cleaned texts were appended as a new column, "Cleaned text"

Following this initial phase, we addressed the challenge of managing the language diversity present in the dataset. To accomplish this, we integrated language detection into our preprocessing pipeline. Leveraging state-of-the-art tools, specifically the “papluca/xlm-roberta-base-language-detection”[1] model from the Transformers library, we accurately identified the language of each comment; this model, capable of identifying over 100 languages, enabled us to classify each comment accurately.

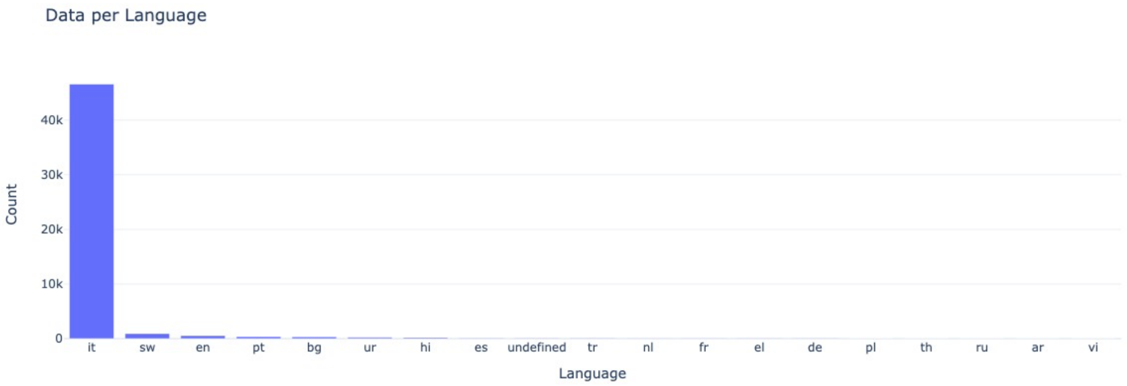

Figure 2, “Data per Language”

[1] You can find more at: https://huggingface.co/papluca/xlm-roberta-base-language-detection

In Figure 2, the distribution plot of languages among the comments is displayed. Italian overwhelmingly dominates the dataset, which prompted our decision to integrate DeepL APIs to facilitate language translation, thus enhancing linguistic diversity and usability. This involved segmenting text to fit model input size limitations and determining the language of each segment, gracefully handling any errors or exceptions. Again, the resulting language labels were appended as a new column, "Language", providing valuable insights into the dataset's linguistic composition. This preprocessing phase laid the groundwork for subsequent NLP analyses, ensuring appropriate handling of language-specific considerations. As a result of this procedure, all non-Italian comments were subsequently translated into Italian.

In the final phase of our preprocessing pipeline, we implemented a critical measure to refine our dataset for sentiment analysis. This involved developing a function capable of grouping comments authored by the same user within a single platform. The aim was to mitigate any potential bias stemming from multiple comments by the same user, thus ensuring a more balanced sentiment analysis outcome. By consolidating approximately 5000 comments, we effectively mitigated the risk of these comments exerting undue influence on the sentiment analysis model's performance and subsequent results. It's important to note that we refrained from grouping usernames across different platforms to maintain the integrity and accuracy of our data preprocessing approach; indeed, it's impossible to ascertain whether the same username across different platforms can be attributed to the same commenter.

2.3. Models

To gain a comprehensive insight into the case study and enhance the conditions for effective sentiment analysis, we opted to implement a four-step pipeline:

1. Keyword extraction
2. Entity recognition
3. Topic modeling
4. Context sensitive sentiment analysis

Firstly, we focused on preparing and processing data by cleaning up the column names, changing the columns with the comments values to analyze Raw Text and Text for higher clarity and easier understanding.

To solve the problem of the presence of stopwords in the model, which could alter the results of numerous tasks such as keyword extraction and topic modeling, we decided to remove stopwords using the library "stopwords" from nltk, used specifically to manage common stopwords libraries in several languages, including Italian.

The set of italian stopwords has been also expanded to include additional common words that are not meaningful for the analysis, ensuring that the text data is clean and relevant for the following analysis.

2.3.1. Keyword extraction

In this section we used the keyword extraction to identify and extract the most relevant keywords from the dataframe, focusing our attention on the most frequent and significant used.

For this task we used a pretrained BERT transformer model, choosing "mrm8488/bert-italian-fined-tuned-squad-v1-it-alfa" which is specifically fine-tuned for Italian.

To initialize the model, we first tokenized the text and extracted keyword features importing the libraries "AutoModel" and "AutoTokenizer" from the transformers package.

Then we performed a series of functions to:

1. Split the text into segments, to handle long texts that exceed the maximum token limit of the BERT.
2. Remove stopwords, to clean and remove common stopwords, as mentioned before.
3. Generate Word Embeddings, using a BERT model to convert words into dense vector representations.
4. Extract keywords, we developed the core of the task using the KMeans clustering algorithm to identify important keywords based on the word embeddings; the keywords are then selected from each cluster based on their distance to the cluster centroids.

Some interesting keywords that we have extracted could be seen in this Wordcloud:

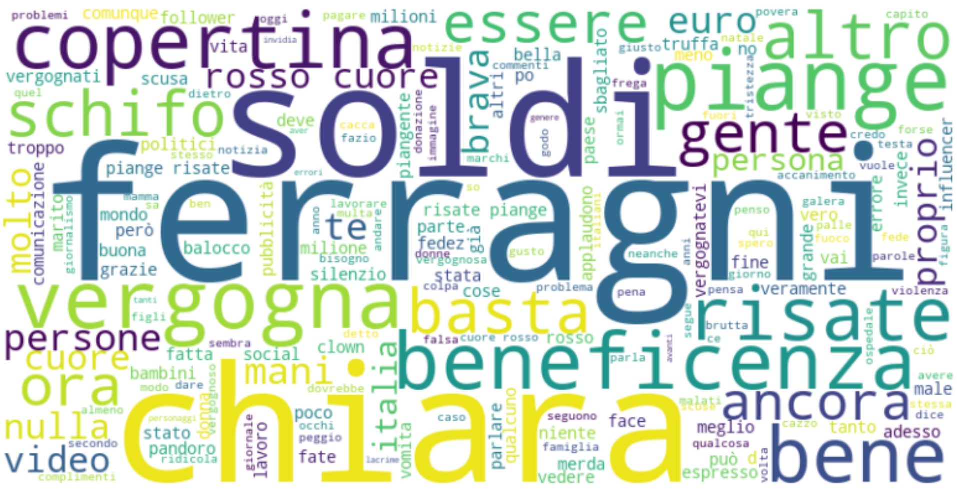

Figure 3: Wordcloud

2.3.2. Topic modeling

Topic modeling represented a crucial step in our analysis, mapping topics and themes within the comments that we scraped. To perform it, we used the BERTopic framework along with the Standord's "Stanza" library for text lemmatization, specifically designed for Italian text.
BERTopic is a flexible and user-friendly topic modeling technique based on BERT embeddings.
We initialized a Stanza pipeline for Italian text lemmatization creating a NLP pipeline that includes tokenzation, multi-word token expansion and, as already told, lemmatization.
Despite our efforts, the topic modeling conducted on the lemmatized text without stopwords yielded unsatisfactory results. However, we have retained the lemmatized text within the dataset for potential future applications.
Then, we initialized and trained the model on the filtered data, having as a output that the model identifies various topics based on the context and relationships between words in the dataset. This process has generated a set of topics along with their respective scores, displayed for a better visualization in a tabular format to show:

- Topic ID, a unique identifier for each topic
- Count, the number of documents associated with each topic.
- Name, the name assigned to each topic.
- Representation, a list of representative words for each topic.
- Representative Docs, some examples of documents associated with each topic.

Moreover, we decided to explore specific topics of interest using the queries that we used also to search data for our dataset, and that were also some of the most important words taken from keywords extraction, such as "Chiara, Ferragni, Beneficienza, Vergogna, Soldi, Truffa".

While processing the topic modeling, we were able to observe that some topics were not relevant to our analysis, not being at all related to the scandal Ferragni-Balocco.
We therefore decided to select and remove them manually from the dataframe in order not to have the need of treating them later in the analysis and to avoid them affecting the performance of the model.
To do this:

1. We worked on an ad hoc dataframe ("df_topic_modeling"), to avoid eventual mistakes
2. We defined a list of unrelated topics by manually examining their keywords and a sample of related texts, and then selecting their IDs.
3. We merged the filtered topic data frame "df_topic_modeling" with the main one "df" to keep relevant original columns and topic IDs; the merged data has been cleaned by filling null values with the most common topics.

4. The notebook filtered out rows where the "Topic" column contains any of the unrelated topic IDs previously listed.

To understand the relation between topics we used an interactive Intertopic Distance Map, that helps in understanding how different topics are related or distinct from each other.
The map shows topics as circles, where the size of each circle represents the prevalence of the topic. The distance between the circles indicates how similar or dissimilar the topics are. Topics closer to each other are more similar, while topics further apart are more distinct.
In the image, the circle shown in red is the topic highlighted by the cursor below, in this case Topic 0.

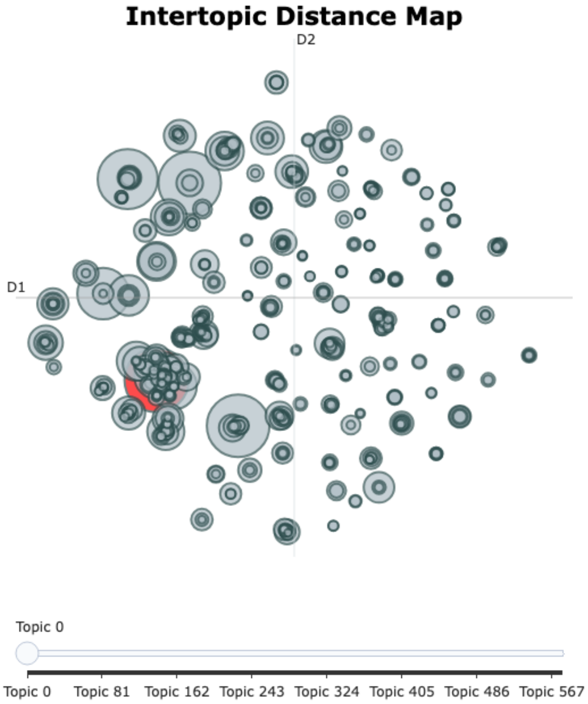

Figure 4: Intertopic Distance Map

As a result of topic modeling 577 different topics were found, with only 10 topics not relevant to our analysis.

2.3.3. Entity recognition

Moreover, we tried to identify entities such as names of people, places, and organizations to understand who the main protagonists are and how they are viewed in the context of the case. This has been made through the Named Entity Recognition (NER), using a pretrained BERT model for Italian text (BERT-IT NER), implementing as done before a transformer model named "nickprock/bert-italian-finetuned-ner", a fine-tuned version of "dbmdz/bert-base-italian-cased".
This transformer model has been used as an initial tokenizer to set up a NER pipeline function, using the aggregation strategy "simple", selecting the most frequent entities with a concatenation made without any additional logic.
The NER is executed without deleting stopwords as their removal could have compromised its performance.
By developing it we noticed that applying the function to the input text, a large part of the comments was not attributed to any entity, and this limited our ability to analyze. In fact, as displayed in the model only 5451 comments out of 38378 have been recognized.
The poor performance of the BERT-IT NER model depends on intrinsic factors such as:

- Quality of input text: the quality of input text can significantly influence the performance of the model. If the text contains many abbreviations, misspellings, unclear sentences, or automatically generated text, the model may have difficulty recognizing entities correctly.

- Text size: the text size may affect the performance of the template. If the text is too short or too long, it may be difficult for the model to correctly identify entities.
- Language variations: Even if the model is trained in Italian, it may not be able to cover all language variants or dialects. If the text contains words or phrases in regional dialects or specialized languages, the model may not be able to recognize them as entities.
- NER model pre-added: The BERT model used may not have been trained on a dataset representative of all input text types. If the text has a specific domain or context that is not covered by model training, it may not be able to correctly recognize entities.

2.3.4. Context Sensitive Sentiment Analysis

Finally we performed the main goal of our project, the sentiment analysis on the data scraped from the web about the Ferragni-Balocco case.
To do so, we used the "osiria/bert-tweet-italian-uncased-sentiment", that is a BERT uncased model for the Italian language.
The model is trained to perform binary sentiment classification (positive vs negative) and it's meant to be used primarily on tweets or other social media posts. A specific feature of this model is that it also considers the context of words before attributing a positive or negative sentiment.
This model was trained on tweets, so it's mainly suitable for general-purpose social media text processing, involving short texts written in a social network style. It might show limitations when it comes to longer and more structured text, or domain-specific text.
In any case, the application of the Osiria model alone is not sufficient to understand the context of a specific comment and thus identify the associated sentiment accurately. Therefore, we have introduced a condition that, if met, modifies the predicted sentiment to improve the accuracy of the prediction in a given context.
We chose to integrate topic modeling to support our sentiment analysis model. Topic modeling allows the model to attribute a preliminary sentiment based on the keywords associated with a particular topic.
The title plays a crucial role: often, a user's comment and thus the associated sentiment are influenced by a psychological bias generated precisely by reading the title. We then categorized titles into "related to a negative context" and "related to a positive or neutral context" using specific keywords selected by us.
Subsequently, we performed sentiment analysis on the cleaned text and made adjustments using the following logic: if the predicted sentiment is positive and at the same time the topic sentiment is positive but the context is negative, then we reversed the prediction to yield a "negative" result. This adjustment allowed us to handle situations such as when a user expresses appreciation for a sanction against Chiara Ferragni. The remaining predicted sentiment was left unchanged.

3. Results

Overall, the sentiment distribution between the comments on the Ferragni-Balocco scandal clearly indicate a predominant negative sentiment, with approximately 30000 negative comments compared to around 5000 positive ones. As it could be seen by the chart below, this distribution highlights the general unfavorable public opinion toward Chiara Ferragni after this scandal.

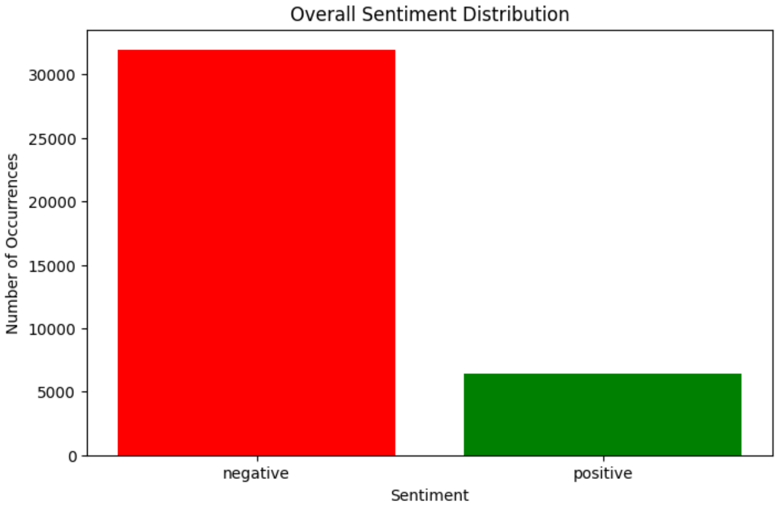

Figure 5: Overall Sentiment Distribution

We also have analyzed the sentiment trends over time to capture the evolution of public opinion. As showed from the chart below, there was distinct periods of extreme negative sentiment, particularly around early January, when the Ferragni-Balocco scandal was published by the media, and March 2024, when the Italian newspaper L'Espresso published on its front page a photo that portrayed Chiara Ferragni dressed up as Joker, with the title " Ferragni Spa: The dark side of Chiara".

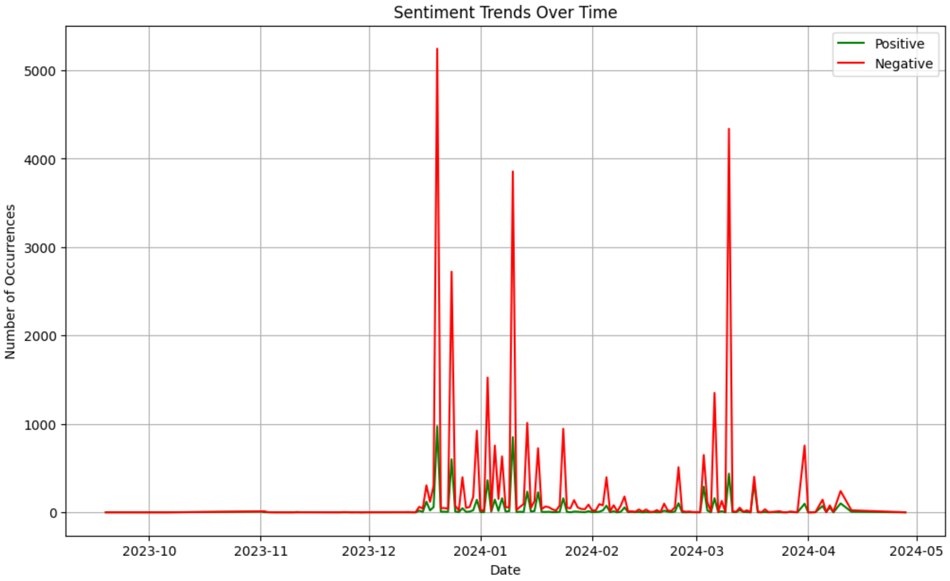

Figure 6: Sentiment Trends Over Time

Considering the 8th of March as a critical date for our analysis, it's possible to see how this event acted as a catalyst for negative public sentiment.

Moreover, by looking at the sentiment across various social media platforms, it's clear that most comments came from Instagram and Facebook, both of which exhibited predominantly negative sentiments. In contrast, YouTube demonstrated a notable proportion of supportive comments, reflecting more balanced perspective among users.

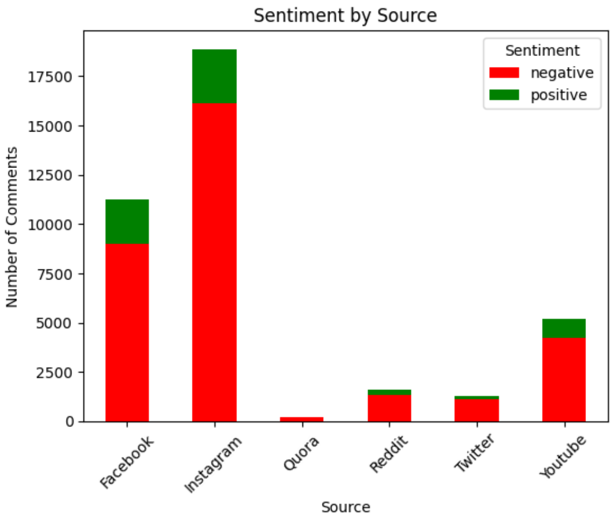

Figure 7: Sentiment by Source

Lastly, we decided to see the distribution of 20 randomly selected topics. By using a heatmap representing the Topic ID, it's possible to see that certain topics, such as "truffa" or "vergogna", were heavily skewed towards negative sentiment, while other topics had a more balanced distribution.

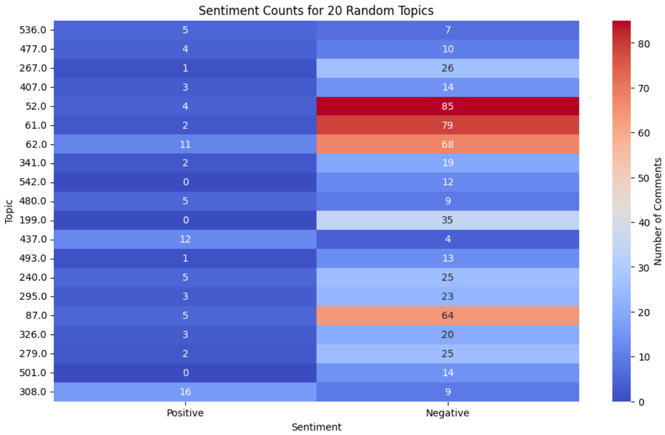

Figure 8: Sentiment Count for 20 Random Topics

4. Discussion and Conclusions

Our sentiment analysis revealed a significant number of negative comments after the great scandal that involved the brand of Chiara Ferragni and the false charity to hospitals for every Balocco Panettone purchased by consumers.

From the data collected, the image of Chiara Ferragni suffered because of this controversy. Some followers have been particularly disappointed or annoyed, highlighting a problem of perception of the authenticity of the influencer. This has led to a decrease in involvement, and consequently a decrease in followers.

The external agents most affected by this scandal were the companies that sponsored the influencer. In fact, the Ferragni-Balocco case has shed light on how the sponsoring companies can be affected by disputes related to their testimonials. Companies have suffered reputational damage due to the bad publicity associated with the case. This has led to a rethinking of sponsorship strategies, with companies now assessing the potential risks associated with influencers more carefully.

After the scandal, Safilo and Coca-Cola decided to stop their collaborations with Chiara Ferragni. Coca-Cola has decided to cancel an upcoming campaign due to consumer concerns related to the scandal. Safilo, who had collaborated with Chiara Ferragni on a line of glasses, also withdrew its sponsorship.

Moreover, several fashion and luxury companies, such as Tod's, Calzedonia and Intimissimi, have only reduced their visibility with Chiara Ferragni.

Other companies are diversifying their portfolios of influencers to reduce the risk associated with a single testimonial, trying to collaborate with personalities who represent values consistent with their brands, avoiding also to get involved in individual scandals.

In conclusion, what can be drawn from this analysis is that the Ferragni-Balocco case has highlighted the potential risks related to the effects of a public scandal on social media:

- For Chiara Ferragni itself, that has observed a decrement of more than 400.000 only from her Instagram account and a great monetary loss in terms of sponsorships. This, however, will not lead to a systematic loss of the interactions and visibility of the contents of Chiara Ferragni because when we talk about people of this caliber, the component of fame always plays an important role.
- For the Sponsors, who as seen had to change their strategies to safeguard their reputation and protect their interests.

Nevertheless, during the "Milan Fashion Week 2024" Chiara Ferragni has generated over 2.5 million dollars in Media Impact Value (MIV) only on Instagram, proving that the influence of collaborations with influencers on social networks is still effective.

Appendix A: Code Description

1. EDA & Pre-Processing: The code starts with importing the necessary libraries and loading the dataset. Make sure that the file "df_with_keywords.csv" is in the same directory as your notebook. If you're using Google Colab, you can upload the file locally and specify its location in read_csv, or upload it to Google Drive and connect it to Colab. At this stage, the dataset is pre-processed for topic modeling and sentiment analysis.
2. Keyword Extraction: In this section, NLP (Natural Language Processing) techniques such as stopwords elimination, lemmatization, and keyword extraction by clustering are used. Make sure that the model and tokenizer are properly imported before running this section.
3. Topic modeling: This part of the code performs topic modeling using the BERTopic package. The notebook presents several graphs and tables to analyze the main themes within the dataset. If you're using Google Colab, make sure you specify the file paths correctly.
4. Context Sensitive Sentiment Analysis: In this section, sentiment analysis is performed using the osiria/bert-tweet-Italian-uncased-sentiment model. The code classifies comments in positive and negative sentiments, taking into account the context. Make sure the model has been downloaded correctly before proceeding.
5. Sentiment analysis for Ferragni-Balocco case: In this part, a specific sentiment analysis is performed on the Ferragni-Balocco case. The notebook uses a file called "df_case-related_topics.csv", which must be generated in the previous sections. If you are using Colab, make sure you link the file path correctly.

Appendix B: Author Contribution

Term	Definition
Data Collection	Gathering raw data through web scraping techniques on social media platforms.
Exploratory Data Analysis (EDA)	Analyzing data sets to summarize their main characteristics, often with visual methods, to identify patterns, anomalies, or check assumptions.
Pre-processing	Preparing and cleaning data for analysis by handling missing data, normalization, encoding categorical variables, etc.
Methodology	Development and design of the overall approach and analytical methods used in the project. This also include: Topic Modelling, Entity Recognition, Keyword Extraction
Study of sentiments	Conducting detailed analysis of sentiments within textual data to understand the predominant emotions and opinions expressed.
Business insights	Deriving actionable and strategic insights from data analysis that have direct implications for business decision-making.
Writing – Reviewing and Editing	Drafting and critical reviewing of the report and code documentation to ensure accuracy, clarity, and comprehensiveness of the content.

Joshua Brauner: Data Collection, Methodology, Business Insights, Study of sentiments, Writing – Reviewing and Editing
Paoloemilio Grande: Exploratory Data Analysis, Pre-processing, Writing – Review and Editing
Simone Luzi: Data Collection, Pre-processing, Methodology, Study of sentiments, Writing – Review and Editing.