Department of Business & Management

MASTER OF SCIENCE IN

DATA SCIENCE & MANAGEMENT

Course: Data Science in Action

SPIRIT

Single Pipeline & Intelligent Reporting for Integrated Tracking

An End-to-End Framework for Brand Monitoring Integrating Semantic Filtering and GPT-Driven Reporting

Supervisor: Alessio Martino

Co-Supervisor: Giuseppe F. Italiano

Candidate: Edoardo D’Onghia

Academic Year: 2023/2024

Table of Contents

Abstract

In a context where a brand’s reputation is constantly shaped by online news, social networks, and industry forums, it is increasingly essential to have tools that can collect, filter, and interpret large volumes of textual data in an integrated and automated way. This thesis presents SPIRIT (Single Pipeline & Intelligent Reporting for Integrated Tracking), an end-to-end system that combines the entire brand monitoring flow into a single platform, from web scraping across heterogeneous sources (news outlets such as Ansa, La Repubblica, Il Giornale, and technical forums like Reddit and Hacker News) to advanced semantic analysis, all the way through to the creation of detailed reports enriched by GPT-generated commentary, and it centers on four main objectives. The first is to offer an integrated pipeline unifying data collection, classification, NLP analysis, and the generation of PDF reports; the second is semantic filtering, which goes beyond simple keyword matching by leveraging Sentence Transformers models to drastically reduce irrelevant content or homonymous references (for example, “Apple” the fruit versus “Apple” the company); the third is user-friendliness, achieved through a Streamlit-based interface that allows the entire process to be configured with natural language prompts, without the need to write any code; and the fourth is reporting automation, which harnesses GPT models to summarize and interpret data in textual form, adding a narrated perspective that is accessible to non-technical users. Experiments demonstrate SPIRIT’s ability to quickly detect changes in public sentiment, identify emerging topics through topic modeling, and provide concise, intuitive summaries, while also highlighting current limitations, including scalability issues, potential “hallucinations” from generative models, and the need for further integrations such as multilingual support, retrieval-augmented generation, and asynchronous processing. Nevertheless, the proposed approach represents a significant step toward increasingly complete, modular, and user-friendly brand monitoring systems, offering substantial potential not only for marketing but also for reputation management and strategic communication.

1. Introduction

1.1 Context and Motivation

Brand reputation monitoring has become indispensable in a world dominated by 24/7 media coverage, social networks, and an ever-growing number of news outlets. Organizations, from global corporations to small businesses, constantly strive to track and manage public perception across digital platforms that range from mainstream newspapers (e.g., Ansa, La Repubblica) to social community forums such as Reddit or specialized tech platforms like Hacker News. By swiftly detecting negative sentiment and controversial trends, brands can identify potential reputational crises early, respond to consumer concerns in a timely manner, and refine their marketing or public relations strategies.

One of the persistent challenges in brand monitoring is the massive volume and variety of online data. Mentions of a given brand might appear in national newspapers, niche tech forums, user-generated social media posts, or aggregator sites. Traditional brand monitoring approaches often focus narrowly, perhaps analyzing only Twitter mentions or employing simplistic keyword matching, leading to partial and sometimes misleading insights. As the digital ecosystem grows, so does the complexity: brands risk missing critical signals if they only monitor a single channel or a single region.

Although numerous commercial tools exist, such as Semrush for brand mentions or specialized social listening dashboards like Brandwatch, these typically focus on partial pipelines (e.g., analyzing sentiment on Twitter alone or pulling data only from publicly accessible Facebook posts). Meanwhile, academic prototypes often concentrate on narrower tasks like topic modeling, sentiment classification, or entity recognition without offering a holistic end-to-end workflow. Both academic and commercial solutions do not always include robust handling of borderline or uncertain data: many rely purely on naive keyword lookups, and relatively few have fully integrated modern large language models to clarify ambiguous items.

This gap motivates the development of SPIRIT, a “Single Pipeline & Intelligent Reporting for Integrated Tracking.” By combining advanced scraping modules, semantic filtering, multiple NLP tasks, and GPT-based commentary, SPIRIT aims to unify the entire brand monitoring journey into one user-friendly system. The overarching goal is to help brand managers, analysts, or marketing teams gather and interpret a wide variety of online data with minimal manual overhead.

1.1.1 Why Sentiment Analysis and Data Gathering Are Still Central in the Age of GenerativeAI

Nowadays, generative AI has captured nearly everyone's attention, from new startups claiming to reinvent industries to big tech companies showcasing ever-more advanced models. It is natural to get excited by the appeal of tools capable of producing complete blog articles, marketing slogans, or even whole design ideas at the click of a button. However, in the middle of this enthusiasm, we risk missing an underlying reality: no AI system, regardless of how creative, can be truly effective unless it is fueled by the right data, and guided by well-formed questions.

From a brand-monitoring standpoint, having "the right data" often starts with scraping. This is where raw information, think user comments, newspaper headlines, forum discussions, gets collected from all sorts of online sources. The scraping process itself is not glamorous, but it is essential because it gives us the raw material for deeper analysis. Without a continuous feed of high-quality text, even the most advanced language model can fall into irrelevance, producing generic or off-target content. In short: scraping is our window to reality. It is how we see what the public is saying about a product, a new feature, or a service glitch, and how we catch emerging topics as they first break.

Once that data is in hand, the next challenge is making sense of it, and that is where *sentiment analysis* becomes a game changer. In marketing and advertising, timing is everything. Catching a wave of negative sentiment early could let a brand fix an issue before it spins out of control, while spotting a sudden burst of positive comments might be the perfect moment for a promotional push. Real-time (or near-real-time) sentiment feedback loops let companies stay agile: pivoting campaign messages, allocating budget toward a popular line of products, or clarifying concerns on social media right when they surface. This sort of responsiveness is not just a nice-to-have; it can decide whether a campaign resonates or falls flat.

Moreover, it is not enough to merely run a sentiment algorithm and produce a percentage saying “60% positive, 40% negative.” The data needs interpretation: in other words, it has to be synthesized into a narrative that managers, marketing staff, or even non-technical executives can understand. That is where a well-crafted pipeline, like what SPIRIT aims to offer, comes into play. You gather data, filter it to ensure it is genuinely about your brand, run analyses (sentiment, topics, entities), and then wrap the insights into a coherent story. At that point, generative AI can be harnessed in a more targeted way: to generate short commentaries that highlight exactly what the marketing or brand team needs to know.

Not to mention that the sheer quantity of online discussion will only keep growing. More news sites, more niche communities, more social media channels, and more user-generated content means that in the future, any brand not equipped to filter and analyze this volume will struggle to make truly informed decisions. Old school, manual approaches, like someone reading a hundred tweets or scanning forum posts by eye, simply cannot keep pace. What is needed are platforms that bring together robust scraping, advanced semantic filtering, and easily digestible reports. SPIRIT, in a modest sense, is an example of how such a platform could look: a framework that tries to unify everything from data collection to AI-based summarization in one place.

While we are not suggesting SPIRIT as the definitive blueprint for the future of data-driven brand intelligence, we do hope it shows where the industry might be heading. By merging real-time data feeds, advanced filters (so we are not flooded with irrelevant chatter), multiple NLP tasks (so we see the “why” behind negative sentiment, not just the “what”), and GPT-driven commentary (so that insights become readable for everyone), SPIRIT hints at a new standard of integrated, user-friendly brand monitoring solutions.

In summary, asking the right questions is at least as important as generating fancy answers, and having the right data to begin with is what makes an AI tool genuinely powerful. Through continuous scraping, real-time sentiment checks, and straightforward interpretations, companies can keep their finger on the pulse of public opinion, using generative AI not merely as a cool gadget but as a guided assistant that helps convert raw data into concrete, impactful actions.

1.2 Problem Statement

The broad question guiding this project is: “How can we collect, filter, and interpret brand-related data in a scalable way, spanning multiple news sources and social channels, and then transform these insights into an actionable, consolidated report?” To answer to this question, we envision a system that must be able to solve four key tasks:

1. **Scrape diverse online sources** (Ansa, La Repubblica, Il Giornale, Hacker News, Reddit, etc.). This involves robust web scraping and API-based retrieval to gather fresh textual content daily.
2. **Filter out irrelevant materials** using advanced semantic embeddings rather than naive keyword lookups, ensuring only brand-pertinent results are retained.
3. **Analyze the textual content** via sentiment classification, topic modeling, and entity extraction, capturing both high-level trends, and more nuanced insights (e.g., frequently mentioned entities or reoccurring discussion themes or most recurrent words).
4. **Summarize results automatically** into a single, streamlined PDF report that integrates visualizations and GPT-based commentary, so business stakeholders can interpret the insights quickly.

Without addressing these steps in a single pipeline, practitioners face a fragmented process often stitching together multiple tools, each of which handles only part of the job. This fragmentation can lead to inconsistencies (different sentiment models, repeated or contradictory analyses) and cause heavy manual overhead in final reporting. SPIRIT attempts to present an end-to-end solution for brand managers seeking a cohesive vantage point.

1.3 Objectives of SPIRIT

The main objectives of SPIRIT can be condensed into four pillars:

- **Pipeline Integration**: Provide a single platform to handle brand mention discovery, data scraping, NLP-based analyses, and result synthesis into a concise PDF file. By doing so, we simplify workflows that typically require separate scraping scripts, different ML libraries, and manual final reporting.
- **Semantic Filtering**: Move beyond raw keyword matching by employing a SentenceTransformer model and similarity-based thresholds. This ensures that tangential or off-topic references to the brand name (e.g., “Apple the fruit” vs. “Apple the company”) do not pollute the dataset.
- **User-Friendliness**: Implement a modern, interactive interface using Streamlit, enabling non-technical users to specify brand keywords, date ranges, or prefer/avoid certain data sources. With a simple prompt, the user can define the entire scraping and analysis logic without writing code.
- **Automated Reporting**: Leverage GPT (Generative Pre-trained Transformer) to produce commentary that highlights key trends, negative spikes, or controversies, thereby reducing manual interpretation overhead for brand managers or analysts. The commentary is integrated with visual elements (charts, top words, etc.) in a final PDF.

Collectively, these objectives address real operational pain points: from excessive data noise and partial coverage to the challenge of explaining trends to non-data-scientists. SPIRIT aspires to let users “click and interpret,” focusing less on engineering complexities and more on strategic decision-making.

1.4 Thesis Structure

Following this introduction, the document is organized as follows: Chapter 2 reviews the academic and industrial literature on brand monitoring systems, sentiment analysis, topic modeling, and commentary based on Large Language Models. Chapter 3 presents a high-level overview of the SPIRIT pipeline, including flow diagrams and module interactions. Chapter 4 examines implementation details for each functional block, ranging from scraping routines to GPT-based reporting logic. Chapter 5 demonstrates how SPIRIT performs with real data, analysing metrics such as sentiment distribution, topic modeling coherence, and overall system overhead. Chapter 6 discusses performance, error cases, GPT "hallucinations," and potential for system scalability. Finally, Chapter 7 summarizes the key takeaways and outlines possible directions for future development.

2. Background and Related Works

2.1.1 Brand Monitoring

Brand monitoring refers to the holistic process of tracking and analysing all public mentions, discussions, and sentiments around a brand across various digital platforms. Historically, many organizations focused on single-source data, most famously Twitter, due to its well-documented API and constant stream of real-time posts. However, the online media ecosystem has evolved: brand-relevant conversations now emerge on mainstream news portals (e.g., Ansa, La Repubblica), community forums like Reddit, niche tech boards (e.g., Hacker News), and a host of social media channels. Many of these mentions might not appear in obvious ways: a product name might be embedded in a user comment on an old article or hidden deep in a Reddit thread discussing product features.

This fragmentation makes it no longer feasible to rely on a single platform or region-specific data. Brand discourse can be scattered, with crucial insights emerging in smaller communities or local papers. Failing to capture these distributed mentions means potentially missing out on emerging controversies or success stories. Hence, modern brand monitoring must expand its scope beyond any single social network, employing tools that collect and unify data from diverse channels. Moreover, because of the varied nature of these sources, some highly formal (newspaper editorials) and others highly informal (user-generated posts), a brand monitoring pipeline requires robust filtering and advanced text analysis.

2.1.2 Generative AI

Generative AI, exemplified by advanced Large Language Models (LLMs) like GPT-4, represents a new paradigm in natural language processing. These models can create coherent human-like text, summarize lengthy documents, respond to complex user queries, and interpret borderline or ambiguous statements. While much public attention currently focuses on chatbots and content generation, Generative AI is increasingly relevant to brand monitoring in two ways. First, it can offer contextual commentary on analytical results. Instead of leaving a marketing team with raw “60% negative sentiment,” an LLM can produce an easy-to-read synopsis, pointing out that negativity spiked on a certain date due to a specific product issue. Second, generative models can adjudicate uncertain or borderline items in a dataset (for example, a mention that might be tangentially related to the brand), deciding if it is in fact relevant. This lets brand managers reduce some of the manual checking that often slows down larger data scraping efforts.

In a world saturated with “generative AI hype,” it is crucial to recognize that LLM-driven commentary does not replace the fundamental tasks of data acquisition, cleaning, and labelling. Instead, generative models augment these tasks by offering interpretive assistance and handling tricky edge cases. For brand monitoring, this synergy is especially vital: LLMs can parse subtle language, detect sarcasm, and interpret coded references to a brand, all while producing concise summaries or concluding remarks about the broader picture.

2.1.3 Sentiment Analysis

Sentiment analysis, sometimes called opinion mining, aims to classify texts (tweets, articles, forum posts, etc.) along a positivity-negativity spectrum. In basic forms, it might simply categorize a post as *Positive* or *Negative*. More advanced systems incorporate nuanced categories (e.g., neutral, anger, sadness, joy), but for brand monitoring, where efficiency and clarity matter, a binary or trinary scheme often suffices.

Older sentiment solutions relied on *rule-based* or *lexicon-based* methods, manually constructing lists of positive words (e.g., “amazing”, “love”) versus negative words (e.g., “hate”, “terrible”). These can quickly falter in real-world contexts where sarcasm, domain-specific slang, or cultural references twist meaning. Today, deep learning and transformer-based approaches are paramount. General-purpose models like BERT, RoBERTa, or domain/language-specific models (e.g., *“MilaNLProc/feel-it-italian-sentiment”* for Italian) leverage contextual embeddings to better capture semantic nuances. A brand mention like “Apple’s new lineup is insanely overpriced" might carry strong negativity that older dictionary methods could miss if the word "insanely" were not recognized as negative in that context.

For brand monitoring, accurate sentiment classification is key to quickly gauging public moods around new releases or crises. Negative sentiment might help isolate problematic topics (e.g., shipping delays, defective products), while positive sentiment helps identify successful campaigns or trending brand advocacy. When integrated with generative AI (as in a pipeline like SPIRIT), these sentiment labels can be automatically expanded into short narratives explaining potential causes or changes over time.

2.1.4 Topic Modeling

Topic modeling is the practice of identifying thematic clusters or latent "topics" within large textual datasets. Latent Dirichlet Allocation (LDA) used to dominate this space, grouping words that frequently co-occur into shared topics. However, LDA's bag-of-words nature often struggles with short texts, local slang, or code-switching, making it less suitable for data that might come from short social media posts or domain-specific subcultures.

A more contemporary approach like BERTopic merges transformer-based embeddings with clustering algorithms, frequently HDBSCAN, to group texts into coherent topics that reflect genuine semantic similarity rather than mere word overlap. In brand monitoring, topic modeling can reveal conversation clusters that might not map cleanly to standard sentiment labels. For instance, a brand might have a "negative topic" around customer support issues and a separate "positive topic" around new design features. By detecting these clusters, brand managers can quickly see which aspects of their offerings attract praise or criticism. Additionally, comparing topics over time (e.g., monthly or weekly windows) can reveal emerging trends or fading controversies.

2.1.5 Relevance Filtering via Embeddings

One major challenge for brand monitoring is the potential for noise or irrelevant references. Simple keyword matching might flood a dataset with false positives, an example often cited is the brand “Apple” vs. references to the fruit “apple”. If the keyword in question is “Musk”, it might confuse references to “Elon Musk” with discussions of, say, “musk cologne”. Such confusion can drastically increase the analyst’s workload, forcing them to sift through countless non-pertinent results.

Embedding-based relevance filtering addresses this by first encoding each piece of text (article, post, comment) into a high-dimensional vector using a SentenceTransformer (e.g., *all-mpnet-base-v2*). The user’s brand prompt (or a set of brand keywords) is similarly encoded. If the similarity (e.g., evaluated via cosine similarity) between a text embedding and the brand embedding is high, the snippet is deemed relevant; if low, it is likely unrelated. A middle zone (where similarity is unclear) can be flagged for a secondary check, often performed by a generative model like GPT, which can read the snippet and explicitly decide whether it is “relevant” or “not relevant.” This hierarchical process drastically cuts down on manual sorting and ensures brand managers do not miss key borderline cases.

2.1.6 Integration for Modern Brand Monitoring

Putting these techniques together, **scraping** from multiple channels, **generative AI** for commentary and uncertain cases, **sentiment analysis** to gauge positivity vs. negativity, **topic modeling** to classify recurring themes, and **embedding-based** filters to ensure relevance, constitutes a next-generation brand monitoring pipeline. Each step addresses a major gap in traditional solutions:

1. **Data Collection**: It starts with systematic scraping or API-based retrieval from newspapers, social media, or specialized forums, ensuring coverage of the full digital footprint around a brand.

2. **Filtering**: The pipeline applies embedding-based checks to filter out noise and keep only brand-pertinent items, alleviating the problem of homonyms, tangential references, or spam.
3. **Sentiment and Topic**: NLP modules automatically reveal user sentiments and highlight the predominant topics driving positivity or negativity.
4. **Generative Summaries and Relevance Checks**: GPT or similar LLMs can handle borderline items during filtering and create commentary that ties the data together into a narrative, smoothing out the edges for non-technical stakeholders.
5. **Reporting**: The system can compile all these results, charts, top words, frequent entities, topic clusters, into coherent textual explanations, drastically reducing the time it takes to finalize comprehensive brand reports.

As the volume of online conversation grows exponentially and consumers spread their feedback across countless niche communities, brand managers need advanced solutions that integrate all these tasks. Traditional brand monitoring often lacked the ability to handle borderline or ambiguous references, forced practitioners to do manual data labelling, or delivered large amount of numbers with little interpretive assistance. By combining generative AI with powerful NLP modules, especially around sentiment analysis, topic modeling, and robust embedding-based checks, modern brand monitoring tools aim to be both more thorough (covering multiple channels effectively) and more user-friendly (offering immediate interpretive narratives and data visualizations).

In this context, **SPIRIT** stands as an illustration of how these concepts can be unified: from initial scraping to final GPT-based commentary. The pipeline demonstrates how each piece (e.g., multilingual sentiment, advanced topic modeling, embeddings for relevance) is not an isolated novelty but part of a coherent framework that addresses real brand-monitoring challenges. This synergy ensures that large volumes of textual data are transformed into actionable insights, allowing companies to promptly address negative feedback, capitalize on positive waves of user excitement, and direct marketing strategies based on authentic, real-time conversation trends.

2.2 Literature Overview

Recent research highlights the growing importance of brand monitoring solutions that incorporate sentiment analytics, topic modeling, and generative AI commentary (Nwohiri & Amaechi, 2022; Barunaha, Prakash & Naresh, 2024; Diaz-Garcia et al., 2022; Dallabetta et al., 2024; Wang et al., 2024). Nwohiri & Amaechi (2022) focus on Twitter-based analysis to capture sentiment shifts in real time, underscoring the urgency of identifying sudden spikes in negativity but revealing the risk of missing important mentions when only one platform is monitored. Barunaha, Prakash & Naresh (2024) likewise emphasize the value of combining sentiment classification with topic detection to better understand the causes behind negative sentiment, although they do not integrate an automated, LLM-driven reporting stage. Other efforts address the challenge of filtering out irrelevant or unreliable data: Diaz-Garcia et al. (2022), for example, introduce a framework aimed at discarding off-topic or low-credibility content, an approach that resonates with SPIRIT's goal of achieving higher-quality brand mention retrieval. Meanwhile, Dallabetta et al. (2024) draw attention to the importance of robust scraping, noting that incomplete or malformed text can compromise the accuracy of subsequent analyses. Finally, Wang et al. (2024) demonstrate how large language models can generate structured commentary on news events, though they also caution against the potential for AI "hallucinations."

Building on these contributions, SPIRIT seeks to integrate robust data ingestion, advanced semantic filtering, multifaceted NLP tasks, and GPT-based commentary into one cohesive environment. By going beyond platform-specific methods and reinforcing its pipeline with embeddings and large language models, SPIRIT aims to reduce manual workload and deliver clearer, more actionable insights, ultimately filling a gap between academic prototypes and purely commercial brand monitoring tools.

2.3 Positioning of SPIRIT

In summary, SPIRIT integrates multiple threads of research: multi-source scraping, semantic filtering, sentiment analysis, topic modeling, and GPT-based commentary. While each of these pieces has been studied in isolation, SPIRIT aims to deliver a cohesive user experience that:

- Minimizes manual overhead through automated “best-of-breed” NLP pipelines.
- Summarizes insights in a final PDF, weaving both charts and language-model commentary into a single deliverable.
- Ensures fewer irrelevant hits appear in the final dataset by combining embeddings with GPT classification for borderline items.

This synergy positions SPIRIT at the intersection of practical brand monitoring workflows and cutting-edge AI research. It neither focuses solely on big-data streaming nor restricts itself to a single social media platform, making it an appealing solution for medium-scale brand intelligence tasks.

3. System Architecture

3.1 High-Level Workflow

Below in *Figure 1* there is a schematic illustration of SPIRIT's architecture:

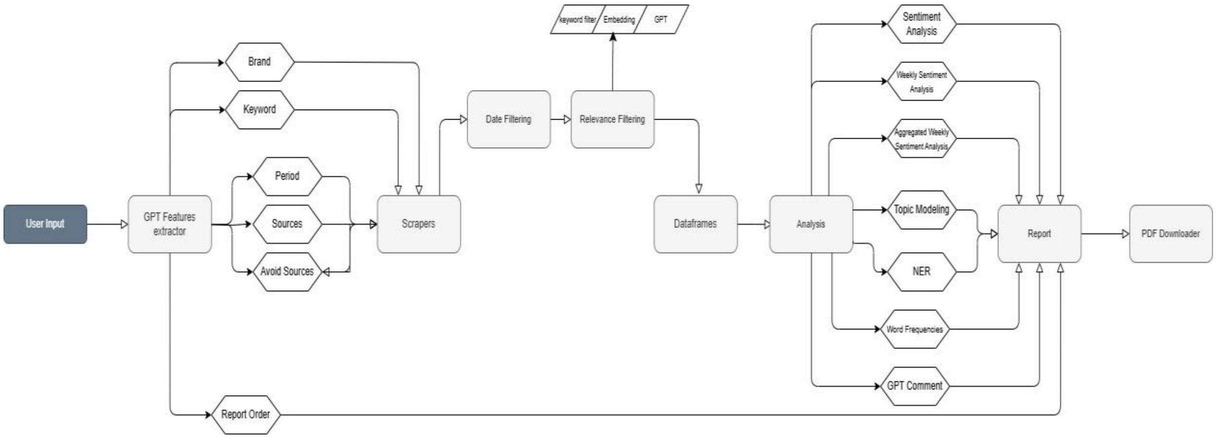


Figure 1. Overview of the web app architecture, from the user input to the pdf output.

1. **User Input**

   An analyst or marketer provides an input prompt describing the brand (e.g., "Apple"), extra keywords ("iPhone 16 Pro"), a date range (start-end), and whether to prioritize or exclude specific news sources and the order of the report (what they want to prioritize as analysis). This step also includes choosing an order of analysis blocks (e.g., time series first, then sentiment, etc.). A major advantage is that the user does not need to manually write code to define each step, rather, they specify a prompt in natural language like:

"I want to monitor Apple from 2023-01-01 to 2023-03-01, particularly the iPhone 16, covering all sources except Reddit. I would like to see immediately the sentiment over time."

2. **GPT Features Extractor**

   Upon receiving the prompt, the system calls a GPT-based parser (parse_prompt_with_gpt) to interpret the user's free-form instructions and extract:

- Brand (single word)
- Keyword or product
- Period (start_date, end_date)
- Sources to prefer (or to avoid)
- Report Order

Each item is captured in a structured JSON format, e.g.:

```
{
  "brand": "Apple",
 "extra_keywords": ["iPhone 16"],
  "start_date": "2023-01-01",
  "end_date": "2023-03-01",
  "prefer_sources": ["Hacker News", "Ansa", "La Repubblica"],
  "avoid_sources": ["Reddit"],
  "report_order":
["time_series","sentiment","topic_modeling","ner","word_frequencies","conclusion"]
}
```

This structured data then configures the subsequent scraping and analysis modules.

3. **Scrapers**
   Based on which sources are “preferred” and not “avoided,” the pipeline runs site-specific scraping functions (Ansa, La Repubblica, Hacker News, Reddit, etc.). These produce raw data for each article or post (title, text, date, etc.), stored in Pandas DataFrames. Immediately after scraping, we apply a basic date filter (only keep items whose dates lie in the user’s chosen period). If the user wants 2023-01-01 to 2023-03-01, older or newer items are dropped.

4. **Relevance Filtering**

   The system then performs a multi-stage funnel to discard irrelevant items:

   i. **Keyword Presence**: If none of the brand keywords appear in (Title + Text), the item is dropped.

   ii. **Embedding Similarity**: We use a SentenceTransformer (e.g., all-mpnet-base-v2) to compute embeddings for the brand query vs. each item's title and text. We define thresholds:

      1. cos_sim_high (say 0.4): items with similarity above this are kept outright.

      2. cos_sim_low (say 0.2): items below this are discarded outright.

      3. Any item in between is labeled "uncertain."

   iii. **GPT Classification for Uncertain Items**: Only borderline items go to GPT with a short prompt: "For each snippet, say 'relevant' or 'not relevant'." If GPT says "relevant," the item is kept, otherwise discarded.

5. **Analysis**

   Once relevance is determined, analysis modules run on the final filtered DataFrame(s). These include:

   - **Sentiment Analysis**: Calls perform_sentiment_analysis (model MilaNLProc/feel-it-italian-sentiment) to label each text as Positive/Negative.

   - **NER**: The function perform_ner applies spaCy's "it_core_news_md" to identify named entities (persons, locations, organizations, etc.), often focusing on negative texts or titles.

   - **Topic Modeling**: perform_topic_modeling uses BERTopic (with HDBSCAN clustering) to discover emergent topics in negative coverage.
   - **Word Frequencies**: For negative texts/titles, the system extracts the top words (excluding brand keywords).

6. **Report (with GPT Commentary) + PDF Download**
   The results from the various analysis steps, sentiment plots, top entities, topic clusters, are compiled into HTML snippets. Meanwhile, GPT (gpt_comment_block) generates short textual commentaries for each stage (e.g., summarizing the time-series trend, highlighting negative spikes, etc.). These snippets, along with images exported from Altair charts, are embedded into a single HTML page. A library (pdfkit) then renders this HTML to PDF.
   - A caching mechanism (cached_gpt_comment_block) avoids repeated GPT calls by hashing the input block data and user prompt, storing the GPT output.
   - Finally, Streamlit provides a "Download PDF" button to retrieve the consolidated brand monitoring report.

3.2 Data Flow

Following the flowchart, data primarily moves from raw HTML into a structured form:

1. **Scraping** yields one Pandas DataFrame per source (title, text, date, etc.). Typically, each scraper function either uses requests+BeautifulSoup (for newspapers) or a specialized API (PRAW for Reddit, Algolia API for Hacker News).
2. **Filtering** transforms or prunes these data, dropping irrelevant entries. Uncertain items go to GPT for a final decision.

3. **Analysis** steps add further columns or produce secondary DataFrames (e.g., weekly_df for weekly time-series sentiment, or a data frame of named entities with counts).
4. **Aggregation**: Partial DataFrames (e.g., from different sources) can be concatenated. Visualizations (Altair/Plotly) are saved locally as PNG for final embedding.
5. **Report Assembly**: The entire pipeline's output is then converted to PDF using pdfkit.

This design keeps data in Pandas for each step, with minimal friction when adding new columns such as Sentiment, Topic, or EntityCount. In essence, the software architecture leverages Python's rich data science ecosystem, layering GPT-based decisions on top to handle borderline classification tasks or commentary generation.

3.3 Caching Strategy

GPT calls can be expensive and time-consuming. The system implements hash-based caching:

- **cached_gpt_comment_block**:
  Each block (e.g., "time_series" or "word_frequencies") plus its relevant data is serialized into JSON and then hashed via MD5.
  - If this hash is in the session cache, we reuse the stored GPT response.
  - Otherwise, we call GPT, store the response under that hash, and return it.

This ensures that if you re-run the pipeline with the same analysis block data, no extra GPT tokens are spent. Given that GPT usage can be costly, especially for borderline classification or large commentary blocks, caching can substantially reduce operating expenses, as well as accelerate repeated analysis runs.

4. Implementation Details

4.1 Technologies and Libraries

SPIRIT’s architecture is built on a variety of Python-based tools that streamline data collection, text analysis, and user interface design. The following sections provide an overview of the main libraries employed, including their purpose, typical usage, and significant observations made during development.

Streamlit

SPIRIT uses Streamlit to create a lightweight web interface for brand monitoring. This library converts Python scripts into interactive web applications with minimal overhead, so non-technical users can easily choose brand keywords, date ranges, or data sources through dropdown menus or text fields. Within SPIRIT, we define pages in the Streamlit session state so that one page prompts for user input, while a separate analysis page shows charts and initiates the final PDF report. To reduce repeated GPT calls, Streamlit’s caching mechanism is combined with an in-code cache. Although Streamlit is an effective solution for smaller or medium-scale projects, it may not be ideal for large-scale environments. If user volume grows significantly, developers may opt for more robust deployment options. After each scraping task or chart creation, SPIRIT updates the Streamlit interface in real time, displaying the latest charts and status messages to keep users informed of the pipeline’s progress.

BeautifulSoup and Requests

For conventional HTML scraping, SPIRIT relies on BeautifulSoup and the requests library to gather data from websites such as Ansa, La Repubblica, or Il Giornale. In this process, SPIRIT constructs and sends an HTTP request for every page, parses the returned HTML, then searches for specific elements that mark relevant information, for example the h2 tag with the class title for article headlines. Depending on the site’s structure, SPIRIT may implement specialized pagination logic, usually fetching a maximum of ten pages for each site, which is generally sufficient for brand monitoring. Website layouts may occasionally change in ways that break the scraper, so SPIRIT handles these cases carefully, returning an empty DataFrame if it cannot recognize the structure. Also, when network conditions lead to request failures or 500 errors, SPIRIT either retries or continues to the next page, ensuring the scraper remains robust.

PRAW (Python Reddit API Wrapper)

To interact with Reddit, SPIRIT uses PRAW to search and retrieve posts from the r/italy subreddit. This tool allows SPIRIT to specify parameters such as sort equal to top and time_filter set to day, week, or another relevant duration, matching the user's date preferences. In order to avoid irrelevant or trivial posts, SPIRIT sets an upvote threshold, for example 25, and excludes posts beyond a certain text length, for example 500 words. PRAW depends on valid API credentials, meaning it requires a client ID and client secret, and it can be constrained by Reddit's rate limits if it processes too many searches too quickly.

Algolia API (Hacker News)

To gather content from Hacker News, SPIRIT uses the Algolia search API by sending requests through the requests library. These requests return JSON data containing story information such as titles, points, and timestamps. To keep scraping times in check, SPIRIT goes through only five to ten pages of results before stopping. The system preserves just the items tagged as story, since job postings or raw comments add little value for brand monitoring. Note that the Algolia API can temporarily limit requests, or offer incomplete results, if the search scope is too broad.

Sentence Transformers (all mpnet base v2)

For the task of relevance filtering, SPIRIT loads the all-mpnet-base-v2 model to encode textual data into dense vector embeddings. The system creates embeddings for each item's title and text, then calculates the cosine similarity against the embeddings for the brand keywords. By doing so, SPIRIT efficiently filters out tangential mentions, for instance apple orchard instead of Apple Inc. Because the model is quite large, often hundreds of megabytes, running it on systems with limited hardware might be slow. Storing embeddings locally can help when re-analyzing the same text multiple times and lessen the load.

BERTopic and HDBSCAN

For topic modeling, SPIRIT combines BERTopic (Grootendorst, 2022) with the HDBSCAN clustering algorithm, enabling the identification of recurring themes within negative textual content. By focusing on items labeled Negative by the sentiment model, SPIRIT highlights user complaints and potential issues tied to the brand. The system displays an overview of the most prominent negative themes, along with top words for each cluster. One should be aware that HDBSCAN sometimes assigns items to an outlier cluster, indicating that they do not fit any main grouping. This behavior might lead to many small, single-document clusters if the dataset is small or too diverse. Additionally, BERTopic's performance can slow down significantly when handling thousands of documents.

spaCy (it_core_news_md)

For Italian text processing, SPIRIT uses spaCy and the it_core_news_md model. This model provides part of speech tagging, dependency parsing, and Named Entity Recognition (NER). After loading the model, SPIRIT runs NER on negative texts or titles, and then collects entity frequencies by label, such as PER, LOC, or ORG. Although this approach is valuable, the model may fail to recognize certain domain-specific entities or slang. In such cases, developers could consider a custom fine-tuned model specialized for brand monitoring. Running spaCy on large datasets can also be time intensive, so SPIRIT limits entity recognition to negative content.

OpenAI GPT

SPIRIT uses GPT 4 for three critical steps. First, GPT extracts parameters from the user's free form input, identifying brand names, additional keywords, date ranges, and which sources to prioritize or ignore. By delivering these elements in JSON format, SPIRIT avoids building separate form fields. Next, GPT classifies borderline items when the embedding-based check cannot confirm whether a text snippet is actually relevant. By grouping these uncertain snippets and sending them to GPT with brief instructions, SPIRIT allows GPT to label them as relevant or not relevant with increased precision compared to simple keyword searches. Lastly, the pipeline alternates analysis and commentary: once each analysis phase concludes, GPT immediately generates its commentary for that phase. For instance, after time series sentiment or topic modeling, GPT produces concise summaries of the most important trends and provides an overall conclusion, turning numeric or chart driven outputs into a more marketing friendly narrative. To achieve these steps, SPIRIT typically sends GPT a system message, describing the JSON fields or borderline classification instructions, followed by user content that is strictly formatted. The main concerns with GPT include API restrictions, the cost of tokens for large scale prompts, and the possibility of hallucinations when GPT invents details not found in the data. To counter that, SPIRIT employs concise bullet style inputs and short answers that remain firmly anchored to the source data.

pdfkit (wkhtmltopdf)

Finally, SPIRIT uses pdfkit (which wraps wkhtmltopdf) to convert HTML and CSS into a downloadable PDF report. After all the analyses have finished, SPIRIT generates HTML that merges GPT’s textual commentary with PNG images, including charts for sentiment or word frequency. Calling pdfkit from string takes this HTML, along with a specified output path and configuration options, and produces a final PDF. The main caveat is that wkhtmltopdf must be installed locally. Also, some CSS or JavaScript based charts may not render perfectly, so SPIRIT exports static images to ensure reliable output.

Why These Tools

Each library covers a different stage of SPIRIT’s brand monitoring pipeline, from scraping text and filtering it to running NLP tasks and creating a final PDF. By combining well known Python libraries, SPIRIT can rely on a wide community of developers who maintain and upgrade these packages. This strategy streamlines the entire flow of brand analysis, letting SPIRIT provide users with an end-to-end solution that effectively discovers insights from raw data and turns them into a polished, readable report.

4.2 Scrapers

SPIRIT includes a dedicated scraper function for each data source. Each scraper returns a Pandas DataFrame with at least three columns: Title, Text, and Date. Below is an overview of common practices:

1. **Ansa**
   - **Period Parameter**
     Ansa's search endpoint can take a parameter like periodo=7 (7 days), 31 (one month), or 365 (one year). SPIRIT automatically picks the minimal parameter needed to cover the user's date range.
   - **Pagination**
     Each page typically returns 12 items; we fetch up to 10 pages if needed. Each page is requested synchronously, with a short pause if necessary to avoid rapid-fire requests.
   - **Data Extraction**
     - We look for <h2 class="title"> for headlines, <div class="text"> for article snippets, and <p class="meta"> for date info.
     - A custom date conversion function handles strings like "25.12.2024, 11:44."
2. **La Repubblica, Il Giornale**
   - **Common Pattern**
     Although the sites differ in layout, the general approach is similar: build a URL containing the user's query, iterate pages, parse HTML for article entries, and store them in a DataFrame.
   - **Unique HTML Structures**

        - *La Repubblica* uses <article> blocks, sometimes burying the date in a <time> attribute.
        - *Il Giornale* might place its article text in <p class="card__abstract">.
    - **Error Handling**
      If a page returns a 404 or the structure changes, the scraper logs a warning and either continues to the next page or returns an empty DataFrame.

3. **Reddit**
    - **PRAW**
      We rely on PRAW for searching *r/italy* with a given query. Because direct date filtering is impossible, the code chooses a “time_filter” (day, week, month, year, all) depending on how many days lie between start_date and end_date.
    - **Noise Reduction**
      SPIRIT discards very short posts (which might just say “hi”) or very long ones (which might be tangential rants). It also requires a minimum upvote threshold, ensuring that extremely niche or spammy posts do not clutter the results.
4. **Hacker News (Algolia)**
    - **Query Structure**
      SPIRIT concatenates keywords into a single string (e.g., "apple iPhone"), calls Algolia’s Hacker News API for up to 10 pages, and accumulates “story” hits.
    - **Data Format**
      The API returns JSON. We parse fields like title, story_text, and created_at. We do not rely on HTML parsing here, Algolia already provides structured data.

5. **Scraper Performance**

   - **Speed**

     Each page request might take 1–3 seconds to complete, depending on network latency. For typical queries (like 10 pages from each source), the process remains manageable (under a minute).

   - **Scalability**

     If brand managers require hundreds of pages daily from multiple sources, we could move to an asynchronous model (e.g., asyncio or concurrency libraries) or schedule overnight scraping tasks.

   - **Caching**

     SPIRIT can be extended to cache previously scraped articles in local JSON or database files, skipping re-download if the same range is requested.

4.3 Relevance Filtering with Embeddings

After scraping, each DataFrame can still contain items only superficially matching the brand name. SPIRIT addresses this mismatch with a funnel designed to progressively refine the data until only truly relevant items remain.

1. **Keyword Presence**
   In the first pass, SPIRIT checks if the user's chosen brand keywords, like "Apple" and "iPhone", appear in the article's title or text. Anything that never mentions them (e.g., "pears and bananas") is immediately removed. This cheap step quickly discards obvious outliers without using more expensive embeddings or GPT calls.

2. **Embedding Similarity**
   For articles that pass the keyword filter, SPIRIT uses an embedding model (e.g., *all-mpnet-base-v2*) to measure semantic closeness to the query. Specifically, it encodes the brand query (e.g., "apple iphone") into a vector and does the same for each article's title and text, then takes the highest of the two cosine similarities as the article's overall relevance score.

   - cos_sim_high (e.g., 0.4): If the article's maximum similarity is above this threshold, it is deemed relevant immediately.

  - cos_sim_low (e.g., 0.2): If the maximum similarity is below this threshold, the article is discarded.
  - Everything between 0.2 and 0.4 is marked uncertain for GPT classification.

3. **GPT Classification**
   For uncertain articles, SPIRIT compiles a concise batch prompt for GPT-4: *"You are a relevance classifier. The user is searching for X. For each text, respond EXACTLY 'relevant' or 'not relevant' in a comma-separated list."* GPT then provides a final verdict on these borderline items, ensuring that only the truly ambiguous cases incur token costs.

Example:
A snippet mentioning "apple orchard expansions in Tuscany" might have a moderate similarity score (say, 0.25). It references "Apple" in a purely agricultural sense, so GPT reads the context and concludes "not relevant," preventing orchard-related articles from polluting a brand-focused dataset.

Tuning

- Adjusting cos_sim_low and cos_sim_high can shift how many articles end up at GPT. A lower cos_sim_high might automatically keep borderline pieces, while skipping GPT classification entirely could leave more mistakes among uncertain cases.
- In practice, using GPT only on the uncertain subset is cost-effective, because highly obvious text (either extremely relevant or extremely off-topic) never triggers GPT.

Concrete Results
In a final test with 30 articles, thresholds of 0.2 (low) and 0.4 (high) produced the following outcome:

- The keyword filter discarded 1 item outright (never mentioned the brand).
- Of the remaining 29, the embedding step marked 3 directly as relevant, 16 as discarded, and 10 as uncertain.
- GPT then confirmed 8 of the 10 uncertain articles as relevant, yielding a final set of 11 relevant documents.

When we checked these decisions against manually assigned labels (i.e., comparing the pipeline relevance with the labelled true relevance), we observed that 0.2 and 0.4 thresholds gave the best trade-off with around 73.3% overall accuracy,

4.4 NLP Modules

Once SPIRIT has narrowed the data to articles that truly focus on the brand, it applies several NLP modules to extract deeper insights:

1. **Sentiment Analysis**
   - **Model**: "MilaNLProc/feel-it-italian-sentiment," specialized for Italian texts, often outperforming baseline models on SENTIPOLC16 (macro-F1 ≈ 0.81, accuracy ≈ 0.84).
   - **Usage**: Each relevant item (title and text) is fed into the model, returning "Positive" or "Negative."
   - **Observations**: Headlines often sound more negative than the article body. By evaluating an entire text, SPIRIT reduces the clickbait effect.
   - **Evaluation**: In the same 30-article experiment where 11 ended up being pipeline-labelled as relevant the model's predicted sentiments agreed with those true labels about 90.9% of the time, underscoring its reliability for brand-related sentiment.
2. **Topic Modeling (BERTopic + HDBSCAN)**
   - **Workflow**:
     1. We focus only on negative items because we want to detect complaint themes.
     2. The model clusters embeddings with HDBSCAN, and then with BERTopic summarizes each cluster with keywords.
   - **Evaluation**: Because there's no ground-truth labelling for topics, we handle them qualitatively. We examine each cluster's top words to see if they form a clear theme, like "shipping delays" pointing to logistics issues, or "financing" hinting at repeated talk of acquisitions. With only 11 final articles, we might see two or three clusters and interpret them by hand, checking if they align with brand context. Although there's no labelled dataset, this human supervisor approach still helps us uncover emerging topics that might otherwise slip by.

3. **Named Entity Recognition (spaCy)**
   - **Focus**: By default, SPIRIT runs NER on negative or final relevant articles to see which people, organizations, or places repeatedly appear.
   - **Outputs**: spaCy's Italian model (it_core_news_md) identifies PER (Person), ORG (Organization), LOC (Location), MISC (Miscellaneous), etc.
   - **Evaluation**: Since we lack a specialized entity-labeled dataset, we rely on the same human supervisor approach used for topic modeling. For instance, does "OpenAI" consistently appear as an organization? Is "Musk" recognized as a person?
4. **Word Frequencies**
   - **Motivation**:
     1. For quick, intuitive scanning, we simply compute the top words (excluding stopwords and brand keywords). This acts as a "lowest common denominator" analysis, simpler than topic modeling but still indicative of frequent complaint words.
   - **Implementation**:
     1. We compile negative texts, tokenize them, remove stopwords (like "di," "la," "che"), and count the results.
     2. The top 10 words can be displayed in a bar chart.

Collectively, these modules let marketing teams see how negative or positive the brand coverage is, what specific themes drive negativity, who is being mentioned in that negativity, and which words the public uses most frequently in complaints.

4.5 GPT-Driven Reporting

Beyond using GPT to classify borderline items in relevance filtering, SPIRIT also employs GPT for generating interpretative text at multiple stages:

1. **Block Summaries**
   - **Time-Series**: After plotting weekly positivity percentages, we call GPT with a JSON snippet describing the data points. GPT returns a short paragraph like: “Positivity remained consistent at around 70%, with a slight dip in early April.”
   - **Sentiment**: GPT might say: “Title sentiment is roughly 60% positive, whereas text sentiment is only 45% positive, indicating more user-level frustrations in the article body.”
   - **Word Frequencies**: GPT can highlight recurring words, explaining how they might connect to brand controversies.
   - **Topic Modeling:** After BERTopic clusters negative articles, SPIRIT can feed GPT the representative keywords for each cluster. GPT then returns a compact explanation like, “Topic 0 centers on ‘shipping’ and ‘logistics’, indicating delivery complaints,” or “Topic 1 references ‘financing’ and ‘investments,’ pointing to repeated discussions of funding.”
   - **NER:** GPT receives a list of frequently detected entities from spaCy (e.g., “Musk,” “OpenAI,” “Rome”) and produces quick insights. This gives the user a human-friendly summary of who or what repeatedly shows up in the brand’s coverage.
2. **Conclusion/Recommendations**
   - **Usage**:
     - A final GPT call can produce a “wrap-up” paragraph with high-level takeaways. For example, “We suggest emphasizing

improved customer service communication around shipping times, given the negative coverage found in your top words.”

- **Prompt Management**:
  - We often pass a carefully structured set of data to GPT to reduce hallucination. We usually specify: “Here are the top words, here are the sentiment percentages, please comment briefly.”
- **Caching**:
  - Because GPT calls can be expensive, SPIRIT uses a caching mechanism. If the data has not changed, we reuse the commentary from previous runs.

Token and Prompt Size

If we pass a large dataset to GPT, we risk hitting token limits. Hence, we compress data into small numeric summaries or top-words lists. For borderline classification, we only send up to 50 uncertain items in each batch to avoid extremely long prompts.

Benefits for End-Users

Rather than manually writing marketing commentary each day, a brand manager can rely on GPT’s short paragraphs to highlight trending issues or potential brand crises, saving time while maintaining consistency.

4.6 PDF Generation

Finally, SPIRIT assembles all textual commentary and images into a single PDF:

1. **HTML Assembly**

   - As each analysis block completes, SPIRIT saves a snippet of HTML. For instance, after generating a sentiment pie chart, the system might store:

     ```
     <h2>Brand - Sentiment</h2>
     <img src="path/to/chart.png" />
     <p>GPT commentary: The brand is stable overall...</p>
     ```

   - These snippets are appended to a list in st.session_state["report_html"]. Once all sources and blocks are processed, the list is concatenated into one big HTML string.

2. **wkhtmltopdf**

   - We call pdfkit.from_string(final_html, "output.pdf", configuration=config, options=options), where final_html is the concatenated HTML.
   - The user clicks “Download PDF,” which streams the resulting PDF file to their browser.

3. **User Flow**

   - The brand manager or analyst, after seeing the charts and commentary in real time, can confirm everything is correct and then download the final PDF.
   - If they want a historical trail, they might re-run SPIRIT periodically and store each PDF with a timestamp (e.g., “Apple_Mar2025.pdf”).

4. **Potential Extensions**
    - Emailing PDFs automatically to a mailing list of stakeholders.
    - Archiving PDF data in a database for longitudinal studies.
    - Using advanced HTML/CSS templates if more polished design is needed (e.g., brand logos, corporate color schemes).

Putting it All Together

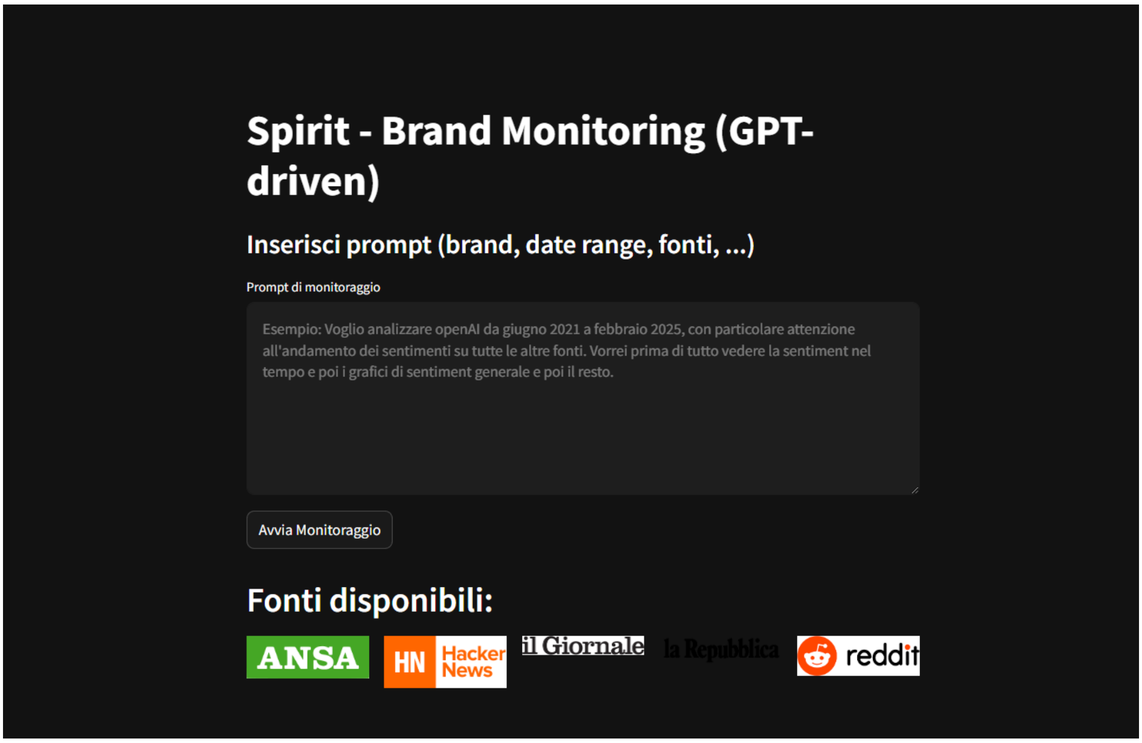
By combining these sections, technology choices, site-specific scrapers, robust filtering with embeddings plus GPT, specialized NLP analyses, GPT-driven commentary, and final PDF generation, SPIRIT offers an end-to-end brand monitoring pipeline. Each component is modular and swappable: if a new site emerges, we just write another scraper; if a new sentiment model outperforms the current one, we replace a function call. This design ethos keeps SPIRIT flexible, maintainable, and ready to adapt to evolving brand monitoring needs.

5. Evaluation and Experimental Results

This section illustrates how SPIRIT performs in a realistic brand-monitoring scenario for “OpenAI,” spanning mid-2021 to early 2025. We begin by introducing the user interface and the typical user flow, then detail the system’s behavior across the multiple stages: Scraping, Relevance Filtering, Sentiment Analysis, Word Frequencies, Topic Modeling, and Named Entity Recognition. Finally, we show how the final textual commentary (GPT-based) aligns with, or occasionally contradicts, real events that shaped public perception around OpenAI in that period.

5.1 Use Case and Interface

When a user accesses the SPIRIT web application, they encounter a straightforward interface:



Spirit - Brand Monitoring (GPT-driven)

Inserisci prompt (brand, date range, fonti, ...)

Prompt di monitoraggio

Esempio: Voglio analizzare openAI da giugno 2021 a febbraio 2025, con particolare attenzione all'andamento dei sentimenti su tutte le altre fonti. Vorrei prima di tutto vedere la sentiment nel tempo e poi i grafici di sentiment generale e poi il resto.

Fonti disponibili:

Figure 2

As we can see in *Figure 2* the interface is in Italian, as the current version of SPIRIT focuses on Italian-language sources. The main elements are:

1. A **Prompt** text area, in which the user types their monitoring request. For instance:

“Voglio analizzare OpenAI da giugno 2021 a febbraio 2025, con particolare attenzione all’andamento dei sentimenti su tutte le fonti. Vorrei prima di tutto vedere la sentiment nel tempo e poi i grafici di sentiment generale.”

In English, that translates roughly to:

"I want to analyze OpenAI from June 2021 to February 2025, paying special attention to the trend in sentiment across all sources. I would first like to see the time-series sentiment, then general sentiment charts."

2. A button "Avvia Monitoraggio" ("Start Monitoring"), which initiates the scraping and analysis pipeline.
3. A display of "Fonti disponibili" ("Available Sources"): e.g., ANSA, Hacker News, Il Giornale, La Repubblica, Reddit.
   By default, SPIRIT can target all these sources, but the user can specify a subset.

Internally, once the user clicks "Avvia Monitoraggio", SPIRIT calls a GPT-based parser to extract:

- **brand** (e.g., "OpenAI")
- **extra keywords** (if present, e.g., "ChatGPT")
- **time interval** (start–end dates)
- **preferred/avoided sources**
- **report order** (time-series sentiment, word frequencies, NER, topic modeling, conclusion, etc.)

5.2 Scraping Validation

SPIRIT scrapes each selected source (ANSA, Il Giornale, La Repubblica, Reddit, Hacker News) for articles/posts referencing "OpenAI," "ChatGPT," or other relevant keywords, filtering by publication date between June 2021 and February 2025.

- **ANSA**: We fetched 47 articles; ~45 contained "OpenAI"-like references. A preliminary embedding check flagged 8 as definitely relevant, 18 uncertain, and 19 discarded for being off-topic (e.g., only referencing "AI" in a non-OpenAI sense). GPT-based classification on the 18 uncertain items kept 12, bringing the final ANSA total to ~20.
- **Il Giornale**: 13 articles found, 5 retained after filtering.
- **La Repubblica**: 6 articles found, 3 retained.
- **Reddit**: 34 posts from r/italy, 9 retained.
- **Hacker News**: 36 hits, and as expected, all 36 were accepted, since Hacker News is a forum about tech news (the brand references were consistently about OpenAI or ChatGPT, so the final dataset for HN is 36).

During various tests performed, as stated before, we found that the thresholds of cos_sim_low = 0.2 and cos_sim_high = 0.4 for the embedding filtering yield the best trade-off for capturing borderline references.

Trade-offs:

- If cos_sim_high is set too high, the system would require a very strong similarity to mark an article as relevant. This can reduce recall because borderline articles, ones that do mention the brand but not strongly, won't pass the strict threshold. As a result, they either become "uncertain" (which then must go to GPT) or get discarded even though they might actually be on-topic.
- If cos_sim_low is set too low, fewer articles get discarded outright. That means even fairly irrelevant ones will end up "uncertain," forcing GPT to classify them. This expands the GPT workload, driving up both cost and latency.

5.3 Sentiment Analysis per Source

After scraping and filtering, each article or post is labeled "Positive" or "Negative" via *MilaNLProc/feel-it-italian-sentiment*. Then, SPIRIT can display weekly time-series sentiment by calculating a weekly mean of sentiment and overall sentiment distributions for "Title" vs. "Text" (respectively the titles and the abstracts of the articles).

Below, in *Figure 3*, will be displayed different results from different sources as a result of the use case prompt.

5.3.1 ANSA Example

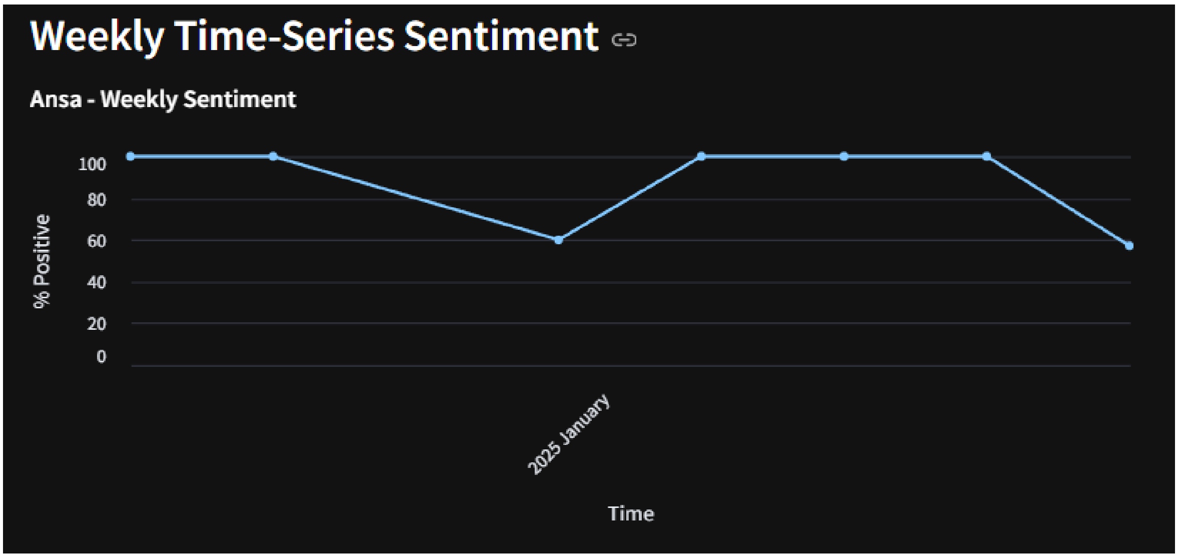


Figure 3

From December 2024 to February 2025, we see high positive percentages in certain weeks (e.g., 100% positivity in early January 2025), though near the end of January 2025 the positivity dips to around 57%. The GPT commentary (in Italian) response was:

“L'analisi dei dati dal Dicembre 2024 a Febbraio 2025 mostra un andamento prevalentemente positivo del sentiment verso OpenAI. La fonte "Ansa" ha pubblicato diverse notizie nel periodo analizzato, con una percentuale di sentiment positivo che ha raggiunto il 100% in diverse settimane. Si nota tuttavia una leggera flessione nell'ultima settimana del 2024 e nell'ultima settimana di Gennaio 2025, dove il sentiment positivo scende al 60% e 57.14% rispettivamente, nonostante il numero di documenti analizzati sia aumentato. Questo potrebbe indicare un cambiamento temporaneo nel tono della copertura mediatica. Nonostante ciò, l'immagine generale rimane positiva, sottolineando un'ampia accettazione e apprezzamento di OpenAI da parte del pubblico e dei media. Questo è un segnale forte per il brand, che può continuare a costruire su questa percezione positiva.”

(Translated to English):

“The data analysis from December 2024 to February 2025 shows a predominantly positive trend in sentiment toward OpenAI. The source “Ansa” published various news stories during the analyzed period, with the percentage of positive sentiment reaching 100% in several weeks. However, there is a slight dip in the last week of 2024 and in the last week of January 2025, where positive sentiment falls to 60% and 57.14% respectively, despite an increase in the number of documents analyzed. This could indicate a temporary shift in the tone of media coverage. Nonetheless, the overall image remains positive, highlighting broad acceptance and appreciation of OpenAI by both the public and the media. This is a strong signal for the brand, which can continue to build upon this positive perception.”

5.3.1.1 Ansa Sentiment Analysis

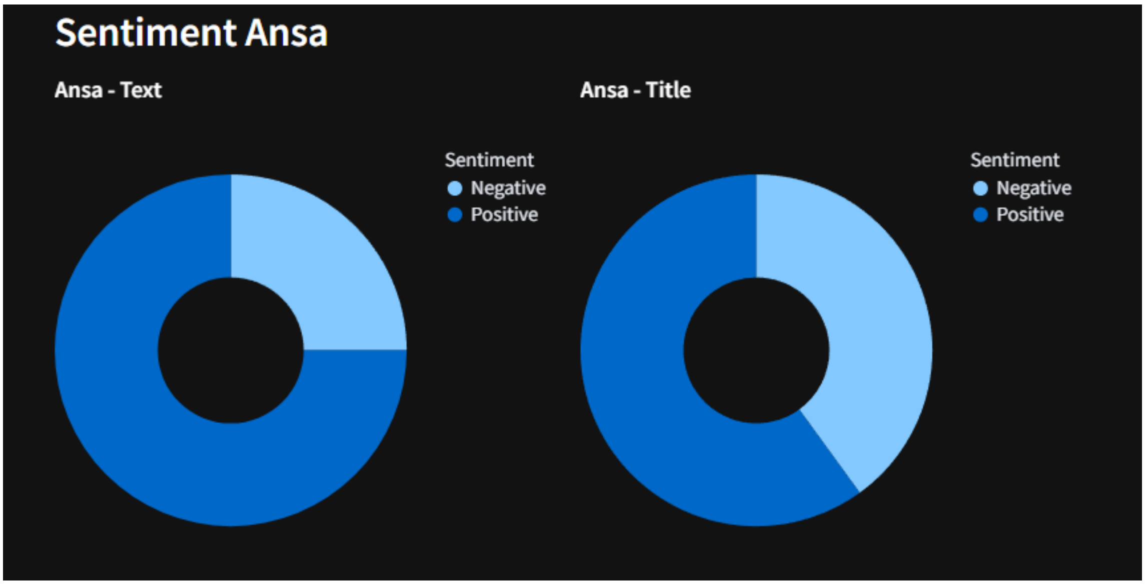


Figure 4

When examining **Titles** vs. **Text** in *Figure 4*, we see ~75% positivity in the textual body, but only ~60% positivity in headlines. GPT's block commentary interprets this gap as a sign that "some negative narratives appear in the headlines," though the general brand coverage is still quite favorable.

5.4 Il Giornale Word Frequencies (Negative)

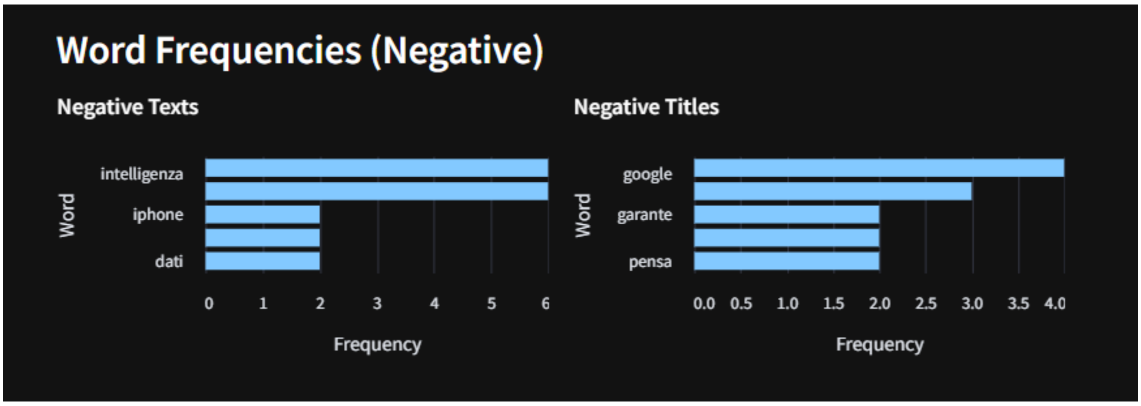


Figure 5

SPIRIT aggregates negative texts/titles from each source and extracts top words, ignoring standard Italian stopwords. The **top-word data** cited above in *Figure 5* comes from the analysis by *Il Giornale*. As seen in the identified texts, the most frequent words are **"intelligenza"** (referring to "intelligenza artificiale"), **"iPhone,"** and **"dati."** This points to a possible partnership between Apple and OpenAI, from which Europe had been excluded due to privacy concerns. Indeed, the term **"garante"** ("garante della privacy") appears in the titles, reinforcing this idea;

meanwhile, **“google”** likely alludes to competition between ChatGPT and Google Gemini.

We can see how GPT interpreted these results:

" The analyzed data indicates a predominance of negative sentiment regarding OpenAI during the observed period. Keywords such as “intelligenza artificiale,” “utenti,” “dati personali,” and “big” were frequently associated with negative contexts, suggesting concerns about privacy and data usage. Specific brands like “Apple” and “iPhone” appear both in negative texts and in titles, hinting at potential tension or competition. It is noteworthy that “Google” is the most frequent term in titles with a negative sentiment, indicating a possible comparison or contrast with this company. Moreover, the presence of words like “sfida” (“challenge”) and “pubblicità” (“advertising”) could point to difficulties or conflicts in these areas. In total, 17 texts and 18 titles with negative sentiment were recorded. These findings call for an in-depth analysis to understand the causes of this sentiment and to develop effective strategies for addressing these issues and improving brand perception."

As we can see, GPT has also cited some additional terms not visible in the graphs. This is because, for better visualization purposes, not all the data provided to GPT is shown in the report. This approach ensures the user has a clear overview, while allowing GPT to include extra details that make the narrative flow more smoothly.

Validation with Real Events

By cross-checking with real developments during Dec 2024–Jan 2025 (see Appendix), we find that ANSA reported on a **15M EUR privacy fine** from the Italian Garante Privacy to OpenAI in late December 2024. This overshadowed otherwise positive news (e.g., a new AI model “o3” specialized in math). Then, in late January 2025, the emergence of “DeepSeek” in China sparked fear that OpenAI’s leadership was threatened. ANSA headlines around Jan 29 mention a potential “war” of AI between OpenAI and Chinese startups. The negativity around privacy controversies and Chinese competition aligns with the 57% positivity drop.

5.5 Topic Modeling and Named Entity Recognition

5.5.1 Topic Modeling

SPIRIT applies **BERTopic + HDBSCAN** to negative items, identifying 2–4 main clusters. Let's analyse the results displayed in *Figure 6*:

1. **Topic 0**: "chatgpt, openai, to" – The first topic, with a count of 5, centers on the keywords *"chatgpt," "openai,"* and *"to"*, indicating that the public's main interest lies in interactions with GPT technology.
2. **Topic 1**: "uomo, sues, embezzled, money" – the second topic, with a count of 2, involves the keywords *"he," "man,"* and *"sues"*, suggesting legal controversies associated with OpenAI. From a brand management perspective, it's essential to closely monitor these emerging themes, as they highlight key areas of public interest and potential challenges.

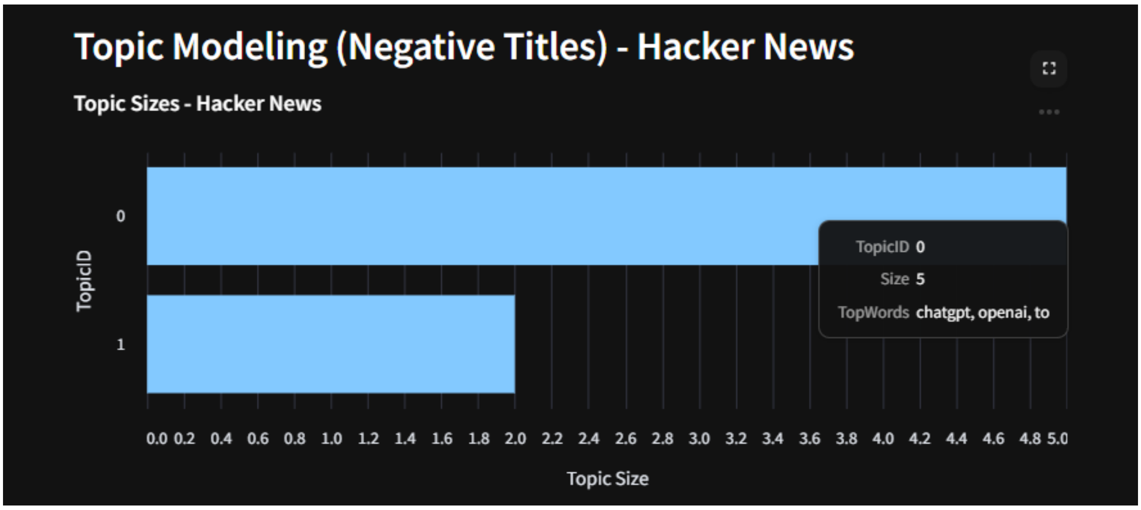


Figure 6

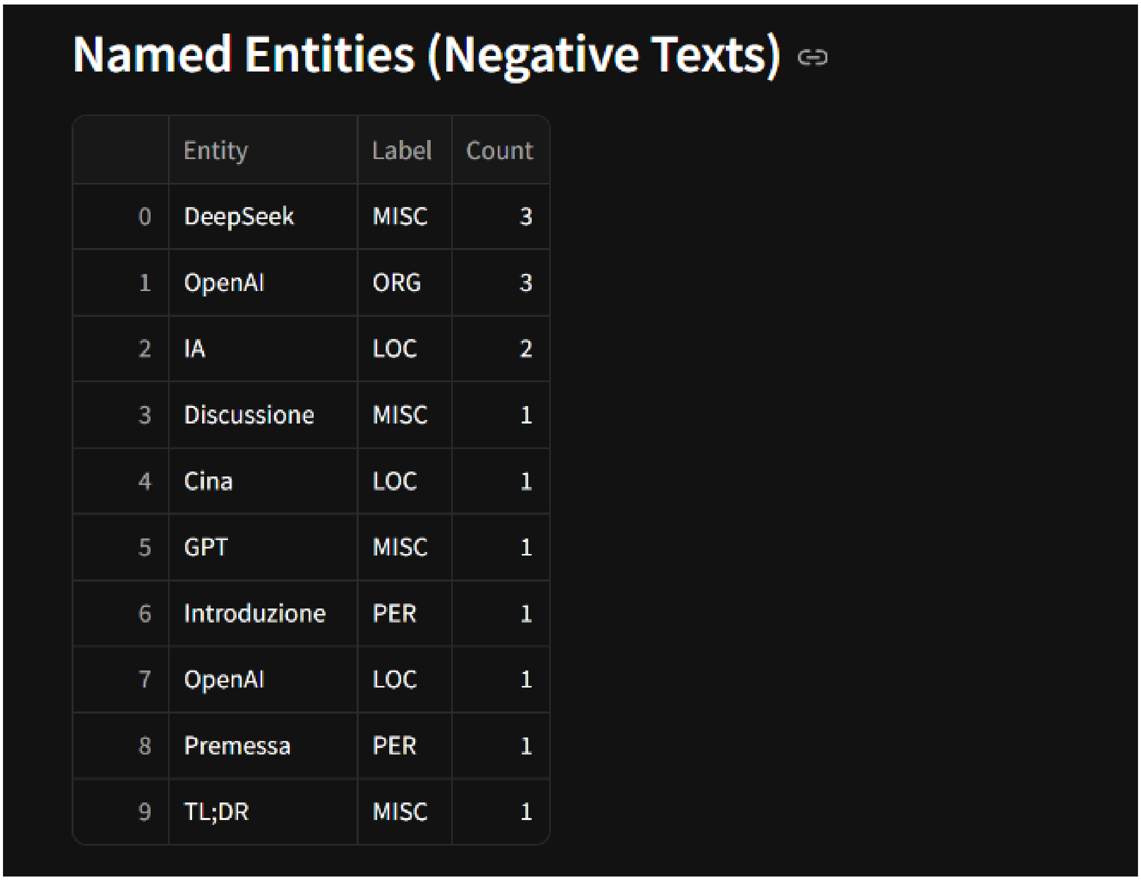
These topics match real legal spats that occurred in 2023–2024, where authors, newspapers, or individuals sued OpenAI for defamation or copyright infringement.

5.5.2 Named Entity Recognition (NER)

The negative corpus often includes repeated references to entities like:

- **OpenAI** (ORG or LOC label, depending on spaCy's guess)
- **DeepSeek** (MISC)
- **IA** or "Cina" (LOC)
- "Garante" (PER) if the text lumps "Garante Privacy" incorrectly
- "GPT" (MISC)

- “Sam Altman” (PER), “Musk” (PER), etc.



Named Entities (Negative Texts)

	Entity	Label	Count
0	DeepSeek	MISC	3
1	OpenAI	ORG	3
2	IA	LOC	2
3	Discussione	MISC	1
4	Cina	LOC	1
5	GPT	MISC	1
6	Introduzione	PER	1
7	OpenAI	LOC	1
8	Premessa	PER	1
9	TL;DR	MISC	1

Figure 7

GPT commentary interpreted these entities in *Figure 7* as “key controversies revolve around OpenAI, DeepSeek, and Chinese references (Cina). Mentions of GPT or Sam Altman (which is not observable from the shown words but is indeed in the dataset) highlight technical or leadership issues.” Real-world news about legal and geopolitical conflicts support that interpretation.

5.6 GPT Commentary and Report Generation

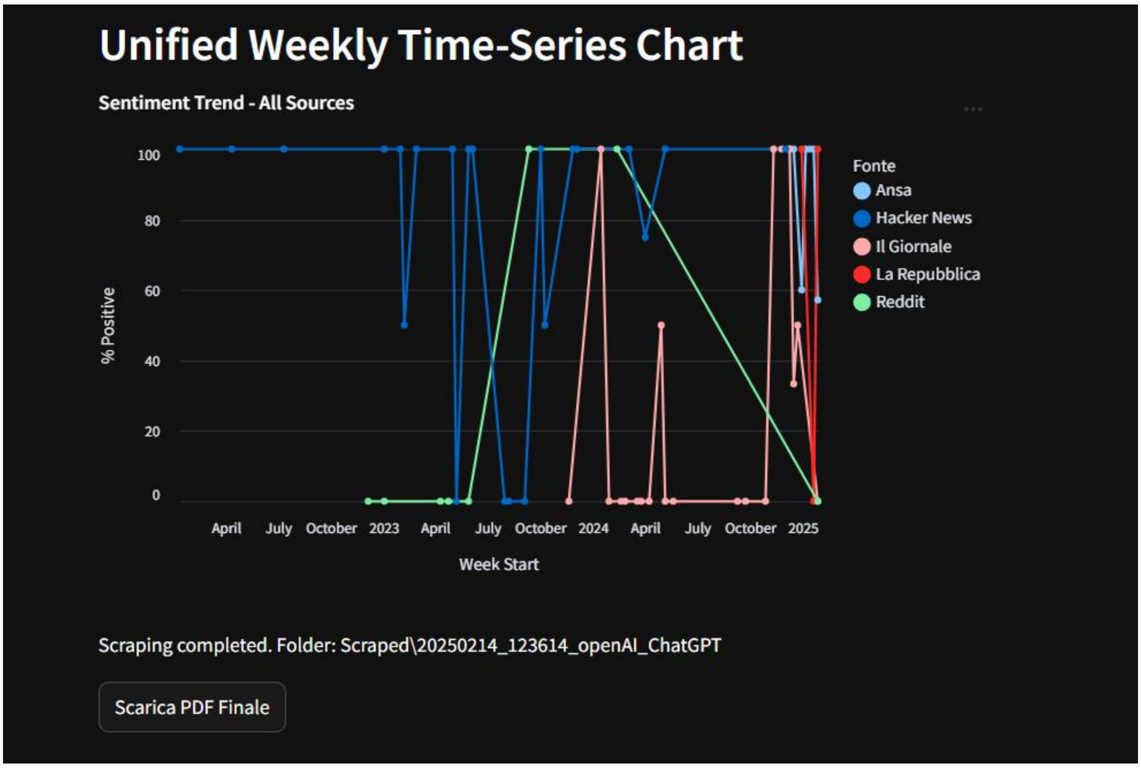
The final step merges each analysis block (time-series, sentiment distributions, word frequencies, NER, topic modeling) into a PDF. GPT generates short paragraphs summarizing each block in Italian. Below is an example from the **conclusion** block regarding Hacker News:

“I dati mostrano un'analisi del sentiment verso OpenAI nel periodo da gennaio 2022 a dicembre 2024, con particolare attenzione alla piattaforma "Hacker News". Nel complesso, l'80.56% dei titoli analizzati ha un sentiment positivo, evidenziando un'immagine generalmente favorevole di OpenAI tra gli utenti. Tuttavia, ci sono stati periodi di sentiment negativo, come evidenziato da alcuni titoli negativi tra cui parole come "sues", "claiming" e "embezzled". Questo

suggerisce che ci sono state controversie o problemi che hanno influenzato l'opinione pubblica su OpenAI durante questo periodo. L'analisi delle entità rilevate mostra una forte associazione con "ChatGPT", suggerendo che questa potrebbe essere una delle principali aree di interesse o di discussione relative a OpenAI. Infine, l'analisi del topic modeling rivela due principali temi di discussione: "chatgpt, openai, to" e "he, man, sues", ulteriormente confermando l'importanza di ChatGPT e le potenziali controversie nell'opinione pubblica su OpenAI.”

(English translation)

“The data presents an analysis of sentiment toward OpenAI from January 2022 to December 2024, with particular focus on the “Hacker News” platform. Overall, 80.56% of the analyzed headlines show a positive sentiment, indicating a generally favorable image of OpenAI among users. However, there were periods of negative sentiment, as evidenced by some negative headlines featuring words like “sues,” “claiming,” and “embezzled.” This suggests that controversies or issues arose that affected public opinion of OpenAI during this period. The entity analysis reveals a strong association with “ChatGPT,” suggesting it may be a main area of interest or discussion related to OpenAI. Finally, topic modeling analysis uncovers two primary themes of discussion: “chatgpt, openai, to” and “he, man, sues,” further confirming both the prominence of ChatGPT and potential controversies surrounding OpenAI in the public eye.”



Scraping completed. Folder: Scraped\20250214_123614_openAI_ChatGPT

Figure 8

At the end of the report, a consolidated time-series chart displays the complete sentiment distribution from all sources over time (*Figure 8*). This offers a broad overview and allows for a clearer understanding of the bigger picture.

Finally, with a single click, the user can instantly download the generated report, ensuring they retain their research and any discoveries made in the process.

5.7 Qualitative Validation Against Real Events

To finalize the evaluation, we cross-check SPIRIT outputs with known historical events:

1. **December 2024 – Privacy Fines Overshadow Positive News**
   The pipeline identifies a sudden spike in negative sentiment toward OpenAI on *ANSA* and *Il Giornale*. This dip coincides with the Italian Garante's decision to fine OpenAI for alleged privacy infractions. The penalty, widely reported by local media, overshadowed otherwise positive announcements (e.g., product updates or new AI features). The sentiment shift is evident in the text analysis: terms like "garante," "sanzione," and "privacy" surge precisely in late December, pulling the overall sentiment down despite ongoing technological advancements from OpenAI.
2. **January 2025 – Competition from "DeepSeek"**
   SPIRIT's output reveals another dip in positivity coinciding with headlines referencing DeepSeek, an emerging Chinese AI model. During this period, *La Repubblica* coverage flips from 100% positive to 0% in a single week, reflecting concern that OpenAI's leadership in AI might be under threat. Negative terms such as "cinese," "rubato," and "concorrenza" appear in the textual data. This abrupt change aligns with real-world events in which DeepSeek's performance, and alleged similarities to GPT, sparked debates on industrial espionage, open-source competition, and potential geopolitical implications.
3. **Reddit – Surges in Positivity (September 2023)**
   The pipeline's analysis detects a pronounced spike in positive sentiment around September 2023 on Reddit. Historically, this period corresponds to the launch of DALL-E 3, along with new "voice and vision" features for ChatGPT. User posts reflect excitement over generating complex images and interacting with AI more naturally, fueling enthusiastic discussions. This perfectly matches SPIRIT's findings: sentiment soared in direct response to tangible feature enhancements and user-friendly improvements,

showcasing how real-time product developments can materially shift brand perception.

4. **Hacker News – Gradual Disillusionment Through 2023**
While initial coverage of ChatGPT in late 2022 was overwhelmingly positive, Hacker News comments reveal mounting skepticism from early to late 2023. This "cooling-off" period emerged as discussions turned to Bing Chat malfunctions (the so-called "Sydney" meltdown), growing legal disputes, and tensions around closed vs. open AI development. SPIRIT's pipeline reflects exactly this trajectory: an early spike of enthusiasm followed by a slow erosion of positivity. By mid-2023, the brand's original "wow" factor on Hacker News started to fade, as discussions about AI regulation, potential performance dips, and open-source alternatives became the main focus.

Overall, these qualitative validations underscore SPIRIT's ability to detect notable sentiment changes in alignment with tangible, real-world events. Whether it is sudden negativity tied to regulatory actions, excitement around product launches, or longer-term disillusionment stemming from legal and ethical challenges, SPIRIT's sentiment metrics consistently track the evolving perception of OpenAI across diverse media sources. Such correlations suggest that the pipeline is a robust tool for gauging brand reputation shifts in real time, capturing both short-lived spikes and more gradual, systemic shifts in public opinion.

5.8 Performance and Overhead

- **Scraping**: 3–5 seconds/page across multiple sources.
- **Filtering**: Running embeddings + GPT classification for borderline items (depending on volume). E.g., 60 uncertain items might take ~10–20 seconds to classify in a single GPT batch.
- **NLP**: Sentiment, NER, and word frequencies are relatively fast (under a few seconds for a few hundred articles).
- **Topic Modeling**: A bit heavier, but workable for up to a few hundred negative items.
- **GPT Comment**: Takes more time to compute since it has to understand the input prompt, analyse and then generate a coherent output (30/40 seconds for each comment)

- **Overall**: End-to-end took ~10 minutes, a fraction of the time a human would spend reading multiple news portals and forum threads.

Note on Computational Environment

All processes described in this work were run on a custom-built personal computer with the following specifications:

- **Motherboard**: TUF Gaming B760M-Plus D4 WiFi
- **CPU**: Intel Core i5-13600KF (14 cores, base 3.5 GHz)
- **RAM**: 32 GB DDR4-3600 (2×16 GB)
- **GPU**: MSI GeForce RTX 3070 Ti Ventus 3X (8 GB GDDR6X)
- **Cooling**: Lian Li Galahad 240 AIO liquid cooler
- **PSU**: Lian Li SP850 850 W SFX (80 Plus Gold)
- **Storage**: 1 TB WD Black SN850 NVMe (PCIe 4.0) SSD
- **Operating System**: Windows 11 Pro 64 bit

This hardware setup was used to scrape data, run text-processing pipelines, and generate GPT-driven commentary for all experiments in this study.

5.9 Summary of Findings

In this use case, the pipeline discovered:

1. **Predominantly positive coverage** for OpenAI, with pockets of negativity triggered by legal controversies or strong competitors (e.g., DeepSeek).
2. **Title negativity** is often higher than text negativity, suggesting that more sensational headlines are used to attract and incentivize the public to read the article.
3. **Key negative words** revolve around “privacy,” “cinese,” “rubato,” “sanzione,” or “accusa,” while “artificiale,” “funzione,” “possibile” appear in negative headlines referencing AI’s limitations or conflicts.
4. **Entities** like “OpenAI,” “DeepSeek,” “Garante,” “ChatGPT,” “IA,” “Cina,” and “Sam Altman” appear frequently in negative contexts, highlighting brand name, competitor, or regulatory authorities.

5. **Two major negative topics** (lawsuits/hallucinations vs. competitor coverage).
6. **Report PDF** final commentary seamlessly stitches these findings into a narrative that business or PR teams can use to respond to brand challenges.

All of these findings were uncovered in a mere ten minutes of automated processing. This efficiency illustrates how SPIRIT significantly reduces the labor-intensive efforts traditionally required for gathering, filtering, and interpreting data across multiple news sites and social platforms. By automating each step, from scraping raw HTML to performing semantic filtering and generating GPT-driven commentary, the system provides a near-real-time snapshot of public sentiment and emergent trends that would normally take hours or days for a human analyst to achieve.

In the following chapter, we will discuss the current limitations of this pipeline, highlighting areas in which the approach could be refined or expanded, such as scaling asynchronous scrapers for higher data volumes or applying advanced retrieval-augmented techniques to reduce hallucinations in GPT commentary. Nonetheless, this case demonstrates a powerful glimpse into the future of marketing research and broader investigative work: gaining actionable insights from disparate, large-scale datasets within minutes, rather than days or weeks, and doing so with minimal human overhead. As digital content continues to proliferate at unprecedented rates, solutions like SPIRIT that harness cutting-edge NLP for streamlined, high-precision analytics will become an integral part of marketing intelligence, competitive research, and organizational decision-making.

6. Discussion and Limitations

6.1 Critical Analysis

SPIRIT’s integrated approach (scraping + filtering + advanced NLP + GPT commentary) drastically simplifies brand monitoring by tying together multiple specialized libraries and methods. End-users gain a one-click pipeline for discovering, analyzing, and summarizing brand mentions. However, some potential pitfalls remain:

- **Scalability**: For thousands of articles daily, synchronous scrapers become time-consuming. GPT calls for borderline documents might scale cost-inefficiently, especially if cos_sim_low is set high, pushing many items to GPT. A robust system would incorporate parallelization, rate-limiting, or more specialized search filters to keep data volumes manageable.
- **GPT Hallucinations**: If the prompts are not carefully structured or if the brand data is ambiguous, GPT might invent spurious “events” or reasons for negativity. We partially mitigate this by passing only short bullet-point data to GPT, but a more advanced approach might be retrieval augmentation, explicitly citing original text passages so GPT stays grounded.
- **Language and Domain Gaps**: The “feel-it-italian-sentiment” model is specialized in Italian. Non-Italian sources or code-switching contexts degrade performance. Similarly, domain-specific slang or references might not be recognized by the embedding filters. A brand like “Apple,” which has an English name but is discussed in Italian contexts, sometimes yields tricky borderline texts.

6.2 The Growing Importance of Brand Monitoring

Over the past decade, the digital landscape has undergone a dramatic transformation. Gone are the days when organizations could focus on just a handful of mainstream channels, perhaps a newspaper or a single social network, to gauge their reputation. Today, brand mentions appear across countless digital ecosystems: local news outlets, community-driven forums such as Reddit, specialized or tech-oriented platforms like Hacker News, and an array of social media sites. Consumers, stakeholders, journalists, and critics collectively shape perceptions of brands around the clock, often in real time.

Simultaneously, the demand for instant insight has escalated. Marketing research is no longer confined to monthly reports or static focus groups. Decision-makers, from public relations staff to top management, need up-to-date analytics on how their product launches, corporate announcements, or controversies are performing across multiple channels. Failure to quickly identify a spike in negative sentiment, or detect the emergence of a damaging rumor, can lead to reputational crises that escalate within hours. In this context, the *speed* of data acquisition and analysis can dictate whether a brand remains resilient or suffers significant harm to its image.

Against this backdrop, SPIRIT (Single Pipeline & Intelligent Reporting for Integrated Tracking) illustrates a possible pathway: a robust approach that merges automated scraping, advanced NLP-based filtering, and AI-generated reporting. By delivering near real-time insights, SPIRIT helps brand managers gain clarity and respond proactively, rather than playing catch-up in a media environment that moves faster than any single human analyst can handle.

6.3 Towards an End-to-End Pipeline

While numerous partial solutions have appeared, ranging from sentiment dashboards for a single social network to classic web-scraping scripts or third-party analytics tools, few systems integrate the entire brand-monitoring process from start to finish. Typically, practitioners must juggle multiple platforms: one for web scraping, another for textual classification, possibly an additional service for topic modeling or entity recognition, and then a separate environment for final data reporting. This fragmentation creates inefficiencies, forces repeated data transformations, and makes it difficult to maintain consistency across each step.

SPIRIT addresses these challenges by knitting together each stage of brand monitoring:

1. **Scraping** from diverse sources (Ansa, Il Giornale, La Repubblica, Reddit, Hacker News, etc.) in Italian, with the potential to add others in the future.
2. **Semantic filtering**, powered by transformer-based embeddings, removing extraneous items and refining borderline cases through GPT classification.
3. **NLP analyses**, including sentiment analysis (Positive/Negative), topic modeling (BERTopic), and named entity recognition (spaCy).
4. **Automated PDF reporting** that calls GPT once again, this time for commentary that weaves these analytical outputs into a coherent narrative.

Such a pipeline not only streamlines operations but underscores the direction in which brand monitoring seems inevitably headed. Eventually, the speed of data creation will exceed any single team's capacity to assess it manually. End-to-end automation, featuring intelligent textual summaries, will enable marketing professionals and communication managers to spend less time on mechanical curation and more on creative strategy or crisis management.

6.4 Data, AI, and the Role of Human Oversight

The global economy has recognized that data has become the "new oil": an invaluable resource fueling industries, marketing strategies, and consumer insights. As a result, the role of the data scientist has grown substantially, becoming one of the most in-demand professions. Yet data grows faster than any single professional can effectively parse on their own, creating a gap that advanced tools must fill.

With the emergence of generative AI (GenAI), many fear that human roles could be replaced entirely by automated systems. Historically, however, each technological revolution, from industrial machines in the 19th century to the arrival of the internet, sparked initial anxieties about job displacement before eventually recalibrating the job market. Humans shifted towards roles requiring creativity, oversight, or deeper strategic thinking that machines could not replicate.

SPIRIT, in this sense, embodies a synergy rather than an outright replacement of human oversight. It harnesses AI at critical junctures, like semantic filtering or commentary generation, but still relies on a user-defined architecture. Software developers and brand managers define thresholds for filtering, decide which sources to scrape, interpret final reports, and guide the system's evolution. This partnership approach ensures that advanced AI modules can handle repetitive or large-scale tasks, while human creativity and strategic sense direct how the outputs should be used.

6.5 Modular, Extensible Architecture

A defining choice behind SPIRIT has been to keep the system modular. Scrapers for each news site or social platform are separated into distinct functions, which can be extended to additional sources as needed. The semantic filtering stage is likewise flexible: we rely on embeddings from a widely recognized transformer model (e.g., SentenceTransformer) and can easily swap out or upgrade this model if a more advanced alternative emerges. The GPT-based classification or commentary steps are similarly organized into function calls that can be updated or replaced with more specialized large language models.

This design ensures that SPIRIT remains future-proof in a rapidly shifting AI landscape. If a newer, more accurate Italian sentiment model becomes available, it can be integrated with minimal friction. If a brand manager wants to monitor niche forums or new aggregator sites that do not yet exist, the code can simply incorporate new scraper modules. By decoupling each block of functionality, SPIRIT exemplifies how brand monitoring solutions can maintain agility in the face of fast-paced AI developments.

6.6 Practical Takeaways and Broader Implications

1. Seamless Flow from Scraping to Reporting:

By orchestrating all stages, data collection, filtering, NLP, and final narrative generation, within a single pipeline, SPIRIT drastically reduces complexity for end-users. Instead of painstakingly linking separate scripts or tools, brand managers can input simple instructions via a web interface and receive a polished PDF with charts and commentary.

2. Speed of Delivery:
In many tests, SPIRIT took around ten minutes from receiving a user's prompt to producing a final PDF report. In a world where crises and controversies can escalate in a matter of hours, or minutes, having a system that can surface a brand's online presence and mood so quickly offers a vital competitive advantage.

3. Global Market Relevance:
While SPIRIT focuses on Italian sources and uses a model specialized in Italian sentiment, the broader concept can be extended to any language or region, assuming an equivalent NLP model is available. This capacity is especially significant for global corporations, whose reputations can vary widely across different countries and languages.

6.7 Comparison to Literature

Compared to Nwohiri & Amaechi (2022) or Barunaha et al. (2024), SPIRIT covers more sources and includes an LLM-based summarization step absent in those works.

This synergy is its strength: from web scraping to final PDF, the user experiences a single integrated pipeline. By contrast, many existing solutions either rely purely on dashboards for real-time analytics or do not incorporate GPT-based textual commentary. The automatic textual commentary aspect is particularly valuable when brand managers desire quick, plain-language insights rather than static charts.

6.8 Ethical and Privacy Considerations

Collecting user-generated content from Reddit or Hacker News can raise concerns about data ownership and user privacy. While we only handle publicly available data, any real-world deployment should respect site-specific terms of service. If the brand monitoring pipeline archives user posts with identifying info, local privacy laws (e.g., GDPR in Europe) might impose constraints. Anonymizing user identities or storing only aggregated results can mitigate privacy risks.

Additionally, GPT commentary can be manipulative if the system is used to create persuasive brand propaganda. One must ensure transparency about AI-generated content, especially in contexts where brand-related narratives might influence consumer decisions. Disclosing that the summary or commentary is AI-generated fosters honesty with stakeholders.

7. Conclusion and Future Work

In an era defined by rapid technological progress and an ever-growing volume of online data, *brand monitoring* has evolved into a mission-critical function for businesses of all sizes. The ability to keep track of public sentiment, competitor activity, emerging trends, and potential reputational risks is no longer merely advantageous: it is essential. As information flows expand exponentially, through global social media channels, online news outlets, niche forums, and specialized platforms, the importance of an automated, end-to-end brand-monitoring framework has become apparent. SPIRIT represents a potential blueprint in this direction, demonstrating how we might unify data scraping, semantic filtering, NLP-driven insights, and AI-generated commentary within one cohesive system.

7.1 Opportunities for Future Enhancements

Despite SPIRIT's capabilities, several avenues remain for further development. These include:

- **Multilingual Coverage:** Expand beyond Italian to track multiple languages, using robust multilingual embeddings.
- **Scalability and Asynchronous Architecture:** Adopt microservices or event-driven pipelines for continuous updates and near-real-time analytics.
- **Retrieval-Augmented Generation (RAG):** Anchor GPT commentary in verified text snippessts (stored in a vector database) to reduce hallucinations.
- **Adaptive Thresholding and Active Learning:** Dynamically adjust relevance thresholds based on user feedback, improving classification over time.
- **User Profiles and Persistent History:** Store queries, brand preferences, and past results for personalized insights and proactive suggestions.
- **Deeper Storytelling and Enhanced UI:** Produce cohesive narratives that connect sentiment, NER, and topic clusters, allowing interactive exploration.
- **Integration with Enterprise Workflows:** Connect SPIRIT to marketing or CRM platforms (e.g., Salesforce, HubSpot) to correlate sentiment trends with campaign data or sales metrics.

This streamlined set of enhancements underscores SPIRIT's potential for delivering richer, more scalable brand intelligence while minimizing manual effort.

7.2 Final Reflections on the Future of Brand Monitoring

As technology accelerates and channels of discourse multiply, brand monitoring's scope and complexity will continue to expand. A single negative mention on a small, highly specialized forum can inflame sentiment across mainstream networks if it resonates strongly with an online community. The capacity to scrape, filter, classify, and interpret an overwhelming tide of content in a matter of minutes is no longer a luxury but a necessity.

Beyond the practical utility, SPIRIT's design also highlights a paradigm shift: generative AI is not purely a novelty tool for superficial tasks, nor is it an existential threat to marketing or data science professionals. Rather, *it is a powerful assistant*, one that can elevate human work, freeing experts from mundane tasks while enabling them to direct the analytics flow at a strategic level. The deep synergy of carefully architected modules (scraping, embedding-based classification, GPT-driven commentary) underscores how AI modules can be integrated responsibly and productively.

In the longer term, as advanced AI models become more ubiquitous, the boundary between data retrieval, analysis, and narrative might blur even further. We may see brand-monitoring systems that automatically adapt to a brand manager's style, or proactively highlight controversies as they emerge, drafting alerts and recommended responses. Tools like SPIRIT could expand well beyond marketing, influencing crisis communications, investor relations, product feedback loops, and policy-making. In essence, *the ultimate promise of AI-driven brand monitoring is to keep humanity in the loop while removing laborious friction from data management*, thereby letting humans focus on empathic reasoning, ethical considerations, and the creative side of shaping a brand's public image.

7.3 Concluding Remarks

SPIRIT exemplifies how diverse technologies, modern scraping, transformer-based embeddings, GPT commentary, and user-friendly UIs, can be woven together to form a cohesive, end-to-end brand-monitoring pipeline. By unifying the entire chain from data collection to interpretative storytelling, it redefines how quickly and thoroughly organizations can gain insight into their brand's reputation and the sentiments swirling around it in the digital sphere.

In a sense, this project is a microcosm of where the broader tech industry is heading: fully integrated systems that harness AI modules at each stage, guided by human creativity and ethical oversight. The expansions discussed here, ranging from multilingual support to retrieval augmentation, are not just incremental improvements, but signals of how brand monitoring will look in the near future: even more dynamic, responsive, and globally oriented. We believe that the SPIRIT framework, or ones akin to it, will become indispensable to any organization wishing to remain agile and connected in an information-saturated world.

Ultimately, the progress achieved by SPIRIT demonstrates that, although the digital landscape grows ever more complex, well-designed AI pipelines can illuminate it with unprecedented clarity, empowering marketers, researchers, and decision-makers to act with confidence, nuance, and speed. By embracing generative AI as a strategic ally, brand managers can more fully engage with the narratives shaping their company's image, respond swiftly to potential crises, and guide public discourse toward constructive outcomes. In doing so, they seize the opportunities presented by the AI revolution rather than becoming victims of its disruptions.

References

Barunaha, A., Prakash, M. R., & Naresh, R. (2023). Real-Time Sentiment Analysis of Social Media Content for Brand Improvement and Topic Tracking. In *6th International Conference on Intelligent Computing (ICIC-6 2023)* (pp. 26-31). Atlantis Press.

Dallabetta, M., Dobberstein, C., Breiding, A., & Akbik, A. (2024). Fundus: A simple-to-use news scraper optimized for high quality extractions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 305–314). Association for Computational Linguistics.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with BERT. arXiv preprint arXiv:2203.00784.

Diaz-Garcia, J. A., Ruiz, M. D., & Martin-Bautista, M. J. (2022). NOFACE: A new framework for irrelevant content filtering in social media according to credibility and expertise. *Expert Systems with Applications*, *208*, 118063.

Nwohiri, A., & Amaechi, C. (2022). Social Media Sentiment Analysis for Brand Monitoring. *Nigerian Journal of Scientific Research*, *21*(1), 110-120.

Wu, Y., Tang, B., Xi, C., Yu, Y., Wang, P., Liu, Y., ... & Yang, M. (2024). Xinyu: An Efficient LLM-based System for Commentary Generation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 6003-6014).