APPLIED AI FOR MANAGERS

FINAL PROJECT REPORT (GROUP 6)

SENTIMENT ANALYSIS AND SUMMARIZATION FOR CALM APP REVIEWS USING BERT & GROQ API (LLAMA3)

Teammates: Thanis Annal Jenifer Kennady (50600953), Keerthi Nagallapati (50598674), Sowfia Syed Abuthaheer (50603739)

1. INTRODUCTION:

Understanding user sentiment is critical to the sustained growth and refinement of mobile wellness applications like **Calm**, which provides a suite of offerings including sleep stories, guided meditations, breathing exercises, and stress-relief content. With a vast and continuously growing volume of user-generated reviews on platforms such as the App Store, Google Play, and third-party forums, manually analyzing this data becomes increasingly impractical, error-prone, and time-intensive.

To address this challenge, our project leverages innovative Natural Language Processing (NLP) techniques combined with transformer-based language models to automate the process of sentiment classification. Specifically, we utilize the GROQ API in conjunction with a BERT-based classifier to interpret and categorize user reviews into two intuitive sentiment categories: **Likes and Dislikes**. Following classification, we also used the GROQ API's LLAMA 3 (8B model) to summarize each review, providing concise insights for stakeholders. We also performed business insight summarization using LLAMA 3 by prompting it as a business analyst to extract two short bullet points - one on what the user liked and one on what the user disliked, each under 10 words.

This automated sentiment engine enables scalable, accurate, and real-time processing of large volumes of textual feedback, transforming raw user comments into structured insights. By doing so, it empowers product managers, UX researchers, and marketing teams to:

- Quickly identify recurring **pain points** and **feature requests**
- Recognize **positive user experiences** and what features drive retention
- Track **user sentiment trends over time** across feature releases or updates
- Make **data-informed decisions** about product roadmaps and experience enhancements

Our solution transforms sentiment analysis from a laborious task into an **actionable, strategic asset** that supports Calm app’s mission to enhance mental well-being through continuous product improvement and responsive user experience design.

2. BUSINESS PROBLEM AND OBJECTIVE:

Problem Statement:

Calm, a leading mobile wellness app, receives a high volume of user reviews across platforms such as the App Store and Google Play. These reviews often contain nuanced emotional feedback, blending appreciation with critique, or expressing sentiment that is context-dependent and subtle.

Traditional sentiment analysis tools typically rely on surface-level keyword detection or basic polarity scoring (positive, negative, neutral). While this approach may work for simple or clearly worded feedback, it falls short when applied to complex or mixed reviews — for example, a user praising the app's meditation content while expressing frustration over subscription pricing.

This limitation creates a gap between available data and actionable insights. Without a more intelligent, context-aware sentiment engine, Calm's product and user experience teams risk missing critical signals embedded within user feedback.

Project Objective:

To overcome these limitations, our project aims to build an automated, high-accuracy sentiment analysis pipeline that enables efficient and insightful review classification. The key objectives are:

- **Leverage a Pre-trained Transformer Model (BERT) via the GROQ API**

  Utilize state-of-the-art NLP capabilities to deeply understand and classify the emotional tone of user reviews. The model is fine-tuned to detect context, tone, and sentiment direction beyond basic keyword matching.
- **Automate Sentiment Tagging for Thousands of Reviews**

  Eliminate the need for manual tagging by automatically classifying reviews as either **"Likes"** or **"Dislikes,"** providing scalable, real-time sentiment insights that can adapt as more reviews are added.
- **Export Structured Output to a CSV File**

  Create a clean, structured dataset that pairs each review with its corresponding sentiment label. This CSV file serves as a foundation for dashboard visualizations, trend analyses, and feature prioritization workflows for Calm's internal teams.
- **Summarize Reviews with GROQ LLAMA 3 (8B Model)**

  After classification, we used LLAMA 3 to generate concise summaries for each review, allowing stakeholders to quickly grasp the core message of long or complex feedback. We also instructed LLAMA 3 to function as a business analyst and generate two short bullet points—highlighting what users liked and disliked under 10 words. Although insightful, these outputs were not always consistent across all entries.

3. DATA COLLECTION & PREPROCESSING:

To build a robust and scalable sentiment analysis pipeline, we began by sourcing user reviews for the Calm app from the publicly available online platform Google Play Store. These platforms contain thousands of user-generated reviews that span a wide range of sentiments, from praise for specific features like sleep stories or guided meditations to concerns over pricing or technical issues.

Once collected, the reviews were saved in a structured format (.csv) to facilitate easy processing and integration with downstream analytical tools. The dataset included the following key fields:

- **Review Content** – The full textual content of each user-submitted review, forming the primary input for sentiment classification.
- **Rating** – The numeric rating (e.g., 1 to 5 stars) assigned by the user. Although this field was not directly used for classification, it was retained to provide contextual depth and could serve as a valuable reference point for future model refinement or correlation analysis.

Preprocessing Steps:

Before submitting the reviews for classification via the GROQ API and BERT model, the dataset underwent a series of preprocessing steps to ensure quality, consistency, and compatibility:

- **Empty Entry Removal**: Any rows containing blank or null review content were discarded to avoid sending irrelevant data through the pipeline.

- **Validation Check**: Each review was checked to confirm it was non-null and contained meaningful content before being passed to the GROQ API for sentiment classification.

These preprocessing steps were critical in ensuring that only clean, relevant, and well-formatted data was used, thus improving the accuracy and efficiency of the sentiment classification process. With a high-quality input dataset, the model was better equipped to detect subtle sentiment cues and generate reliable Like/Dislike labels.

4. SENTIMENT CLASSIFICATION WITH BERT & GROQ API:

To automate sentiment tagging of Calm app reviews, we integrated GROQ's high-speed inference engine with a BERT (Bidirectional Encoder Representations from Transformers) model—one of the most effective transformer architectures for contextual natural language understanding.

This integration allowed us to harness the power of deep contextual embedding while maintaining high performance and low latency, making it ideal for scalable, near real-time sentiment analysis.

Implementation Highlights:

- **Sequential Processing via API Calls:**

  Each user review from the cleaned dataset was passed individually to the GROQ API endpoint using a simple HTTP request. This ensured consistent formatting and allowed for fine-grained control over the classification process.
- **JSON-Based Response Handling:**

  The response from the API came in a structured JSON format, containing a sentiment prediction under the key ['output']['label']. The value returned was either:

    - "POSITIVE" – indicating the review expressed a favorable sentiment
    - "NEGATIVE" – indicating the review conveyed dissatisfaction or critique

- **Mapped Output into Intuitive Columns:**

  To simplify downstream analysis and visualization, we translated the raw prediction labels into two binary indicator columns within our DataFrame:

    - **Likes** – Marked with a 1 for reviews classified as "POSITIVE", 0 otherwise
    - **Dislikes** – Marked with a 1 for reviews classified as "NEGATIVE", 0 otherwise

- **Summarization via LLAMA 3 (8B):**

  Each review was also passed to GROQ's LLAMA 3 model to generate a concise, human-readable summary. This summary was stored in a third new column titled 'Professional Summary', enhancing the dataset's interpretability for dashboards and quick insights. We also added two more columns, 'Likes' and 'Dislikes', based on a business insight summarization prompt instructing LLAMA 3 to function as a business analyst and extract two bullet points under 10 words. While useful, the outputs were not always consistent for every observation.

This approach enabled a fully automated sentiment classification and summarization pipeline with no manual intervention, allowing us to tag and distill thousands of Calm reviews accurately and efficiently. The resulting labeled and summarized dataset was then used for further dashboard visualizations, insight generation, and product feedback analysis.

5. OUTPUT & DATA ENRICHMENT:

- Once sentiment classification was completed via the GROQ API and BERT model, we transitioned to the final stage of our pipeline: enriching and exporting the dataset for downstream analysis and visualization.

Enrichment of the Original Dataset:

- The output from the sentiment classification was integrated directly into the original review dataset:
- **Two New Columns Added**:
  - **Likes** – This column was assigned a value of 1 for reviews predicted as *POSITIVE* and 0 otherwise.
  - **Dislikes** – This column was assigned a value of 1 for reviews predicted as *NEGATIVE* and 0 otherwise.
- **Summary** – A short, general-purpose summary generated by LLAMA 3.
- **Likes** – Bullet point of what the user liked (under 10 words).
- **Dislikes**– Bullet point of what the user disliked (under 10 words).

These summaries were produced using a role-prompted business analyst persona, but some entries may vary in clarity or completeness. This multi-column structure allowed for clear binary classification and interpretive insights, simplifying filtering, aggregation, and visualization in subsequent analytical steps.

Exporting the Final Dataset:

- The enriched dataset was saved in CSV format for easy integration with analytics tools such as Tableau, Excel, or Python-based visualization libraries.

Applications of the Enriched Dataset:

The processed dataset, now containing both original review content and corresponding sentiment classifications, opens the door for a wide range of data exploration and business intelligence use cases:

- **Dashboard Development:**

  The dataset can be loaded into tools like **Streamlit, Tableau, or Power BI** to build dynamic dashboards showing trends over time, sentiment breakdowns by feature, or regional differences in user feedback.

- **Frequency & Pattern Analysis:**

  By counting the number of Likes and Dislikes, product teams can identify which app features receive the most praise or criticism and monitor how these sentiments evolve with each app update.

- **Visual Summaries:**

  Using natural language techniques and visualization tools, we can generate:

  - **Word clouds** highlighting common praise and complaints
  - **Bar charts** comparing sentiment across user segments
  - **Pie charts or heatmaps** showing sentiment distribution by rating or time period

This enriched file serves as an asset for Calm’s product managers, UX designers, and marketing teams seeking to make data-driven decisions grounded in real user feedback.

6. RESULTS & INSIGHTS:

After completing the sentiment classification and enriching the dataset, we performed a comprehensive analysis of the labeled results to assess both **quantitative trends** and **qualitative accuracy**.

Key Observations from the Final Dataset:

- **Predominantly Positive Sentiment:**

  A large proportion of reviews were classified as **positive**, indicating high user satisfaction with Calm's core features. Users frequently praised:

  - The **soothing voices** used in meditation and sleep stories
  - The variety and effectiveness of **sleep aids**
  - The app's overall contribution to **stress reduction and relaxation**

- **Notable Negative Feedback Themes:**

  Among the reviews classified as **negative**, recurring themes included:

  - **Technical issues**, such as bugs, app crashes, or slow performance
  - **Account and login difficulties**, especially after app updates
  - **Subscription-related concerns**, including pricing, auto-renewal confusion, and limited free content

These insights help Calm's product and customer support teams prioritize high-impact areas for improvement.

Manual Evaluation of Model Accuracy:

- ○ To validate the effectiveness of the sentiment classification pipeline, we manually reviewed a **representative sample** of the labeled data. The results highlighted the strength of using a **BERT-based model via the GROQ API** for nuanced language understanding.

Highlights of Model Performance:

- **Mixed Reviews:**

  BERT was able to manage reviews containing **both praise and criticism**, accurately capturing the overall tone by weighing contextual clues rather than relying on isolated keywords.

  Example:

  "Love the meditation sessions, but the app keeps freezing. Hope they fix it soon."

  → Correctly tagged as **Negative**, reflecting user frustration despite a positive mention.

- **Subtle and Emotive Language:**

  The model effectively recognized **indirect expressions of dissatisfaction**, such as sarcasm, disappointment, or frustration without overt negativity.

  This capability surpasses traditional sentiment scoring methods, which often misinterpret such nuances.

  Example:

  "Used to be my favorite app, but now I can't even log in."

  → Correctly classified as **Negative**, despite initial positive sentiment.

- **Summarization Adds Interpretability:**

  In addition to sentiment tags, the LLAMA 3-generated summaries made it easier for stakeholders to quickly review core feedback themes without reading lengthy reviews. This summarization layer enhanced accessibility and interpretability for business users. The business insight-style summaries ('Likes' and 'Dislikes') offered clear, concise breakdowns of user sentiment when successful, though consistency varied across entries.

7. STREAMLIT INTEGRATION:

Although originally developed for internal analysis, the application pipeline has now been successfully deployed on Streamlit Cloud, making it accessible to stakeholders through a user-friendly web interface. The pipeline is fully compatible with Streamlit and provides a solid foundation for future development and scalability.

Key Capabilities and Future Enhancements:

- **Real-Time Review Upload**: Users can upload new customer reviews on the fly, enabling dynamic updates to the sentiment analysis pipeline without needing manual data refreshes.
- **Sentiment Display Dashboards**: Interactive dashboards visualize sentiment trends over time, categorizing feedback as positive, negative, or neutral. These insights can help teams understand customer emotions across product updates or campaigns.
- **Likes/Dislikes Summary with Visualizations**: Aggregated statistics and charts offer a snapshot of overall user satisfaction, helping identify which features are well-received and which need improvement.

Value Proposition:

This app provides Calm's internal teams, product managers, and social media strategists with a powerful tool to monitor and interpret customer sentiment in real time. By centralizing feedback insights in a visual and interactive format—including both sentiment tags and LLAMA 3-generated summaries, the tool supports data-driven decisions to enhance user experience, content strategy, and customer engagement initiatives.

8. CHALLENGES:

- **API Limitations**: GROQ's free tier and documentation were minimal, requiring custom parsing logic.
- **Sarcasm/Irony**: Some complex emotional reviews were misclassified (a common limitation of all NLP models).
- **Latency**: Review-by-review submission was time-intensive for large datasets.
- **Model Cost and Throughput Constraints**: Running both sentiment and summary inference sequentially increased compute time and resource usage, requiring optimization in batching strategies.
- **Prompt Sensitivity in Business Summaries**: The prompt-based summarization approach using LLAMA 3 for bullet-point Likes/Dislikes sometimes produced inconsistent outputs across similar inputs. This may require post-processing or retraining for stability.

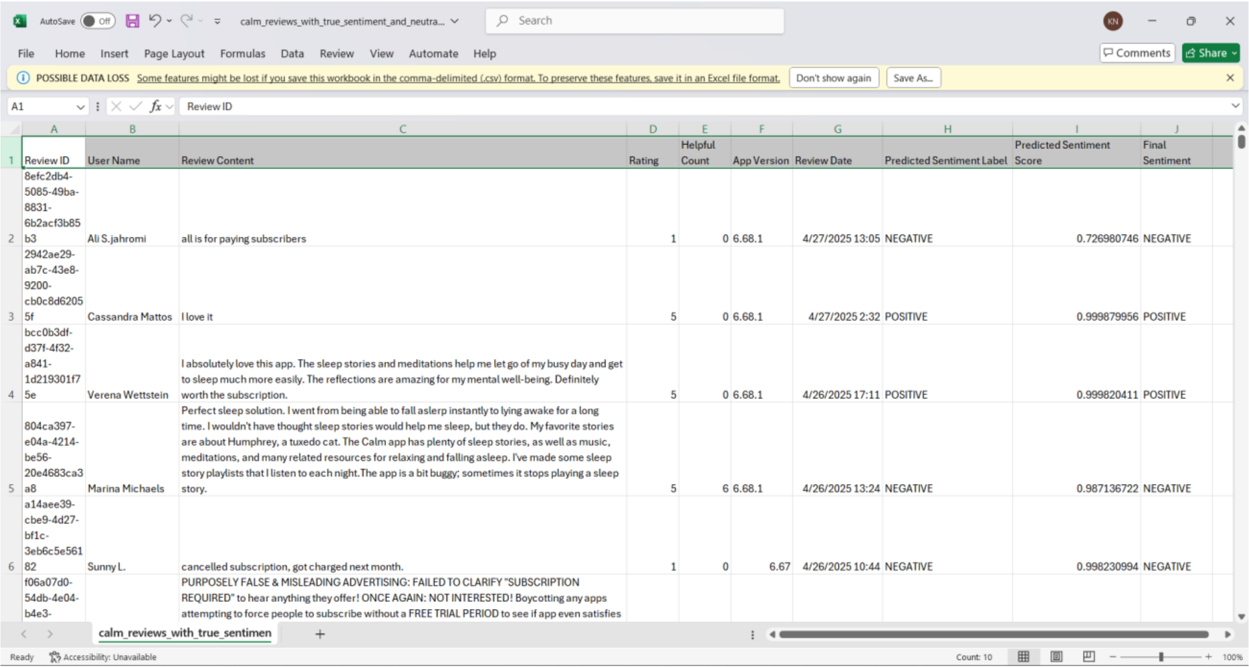
9. CONCLUSION:

This project demonstrated how transformer-based sentiment analysis can be applied to user-generated app reviews to uncover product insights. Using the GROQ API with BERT, we automated sentiment tagging and prepared a clean and enriched dataset for further analysis. In addition, the integration of the GROQ LLAMA 3 (8B model) enabled us to automatically generate both general summaries and business-style bullet points summarizing Likes and Dislikes.

In the future, we aim to:

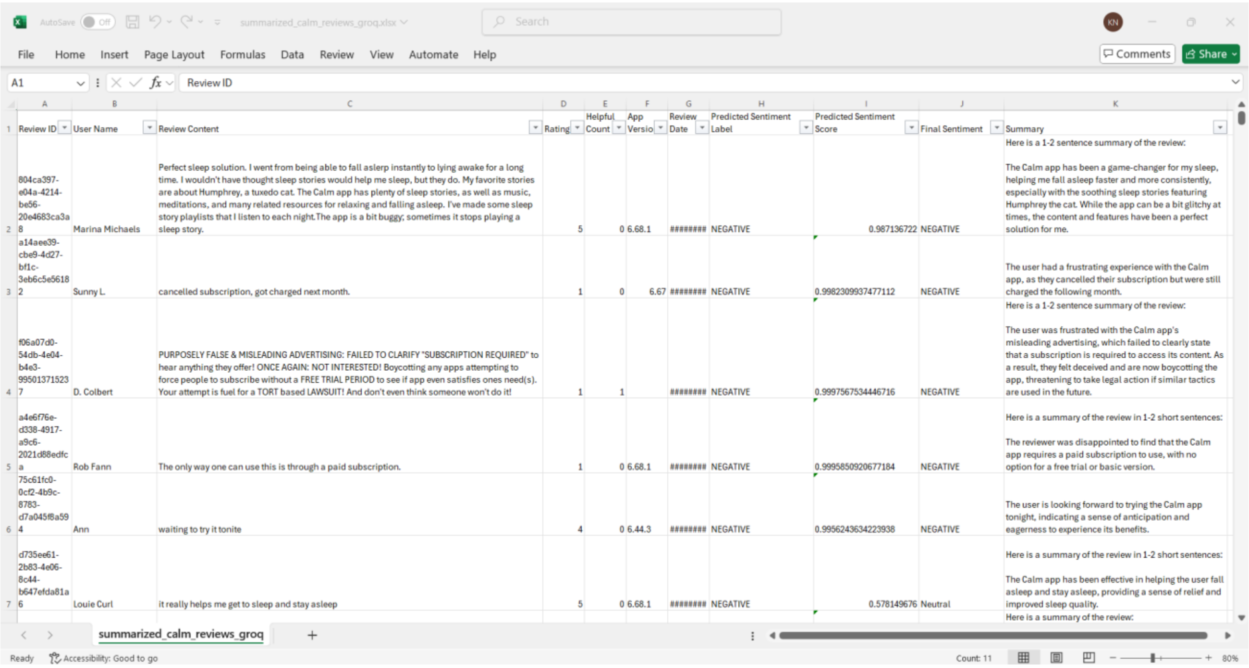
- Incorporate aspect-based sentiment classification (e.g., separate sentiment for UI, content, pricing)
- Support multi-language reviews
- Enable summarization-based clustering for thematic grouping of reviews
- Improve consistency and clarity of business-style summary prompts.

10. APPENDIX:



Review ID	User Name	Review Content	Rating	Helpful Count	App Version	Review Date	Predicted Sentiment Label	Predicted Sentiment Score	Final Sentiment
8efc2db4-5085-49ba-8831-6b2acf3b85b3	Ali S.jahromi	all is for paying subscribers	1	0	6.68.1	4/27/2025 13:05	NEGATIVE	0.726980746	NEGATIVE
2942ae29-ab7c-43e8-9200-cb0c8d62055f	Cassandra Mattos	I love it	5	0	6.68.1	4/27/2025 2:32	POSITIVE	0.999879956	POSITIVE
bcc0b3df-d37f-4f32-a841-1d219301f75e	Verena Wettstein	I absolutely love this app. The sleep stories and meditations help me let go of my busy day and get to sleep much more easily. The reflections are amazing for my mental well-being. Definitely worth the subscription.	5	0	6.68.1	4/26/2025 17:11	POSITIVE	0.999820411	POSITIVE
804ca397-e04a-4214-be56-20e4683ca3a8	Marina Michaels	Perfect sleep solution. I went from being able to fall aslerp instantly to lying awake for a long time. I wouldn't have thought sleep stories would help me sleep, but they do. My favorite stories are about Humphrey, a tuxedo cat. The Calm app has plenty of sleep stories, as well as music, meditations, and many related resources for relaxing and falling asleep. I've made some sleep story playlists that I listen to each night.The app is a bit buggy; sometimes it stops playing a sleep story.	5	6	6.68.1	4/26/2025 13:24	NEGATIVE	0.987136722	NEGATIVE
a14aee39-cbe9-4d27-bf1c-3eb6c5e56182	Sunny L.	cancelled subscription, got charged next month.	1	0	6.67	4/26/2025 10:44	NEGATIVE	0.998230994	NEGATIVE
f06a07d0-54db-4e04-b4e3-		PURPOSELY FALSE & MISLEADING ADVERTISING: FAILED TO CLARIFY "SUBSCRIPTION REQUIRED" to hear anything they offer! ONCE AGAIN: NOT INTERESTED! Boycotting any apps attempting to force people to subscribe without a FREE TRIAL PERIOD to see if app even satisfies							

Figure 1: CSV with positive or negative sentiment prediction of the user review



Review ID	User Name	Review Content	Rating	Helpful Count	App Versio	Review Date	Predicted Sentiment Label	Predicted Sentiment Score	Final Sentiment	Summary
804ca397-e04a-4214-be56-20e4683ca3a8	Marina Michaels	Perfect sleep solution. I went from being able to fall aslerp instantly to lying awake for a long time. I wouldn't have thought sleep stories would help me sleep, but they do. My favorite stories are about Humphrey, a tuxedo cat. The Calm app has plenty of sleep stories, as well as music, meditations, and many related resources for relaxing and falling asleep. I've made some sleep story playlists that I listen to each night.The app is a bit buggy; sometimes it stops playing a sleep story.	5	0	6.68.1	########	NEGATIVE	0.987136722	NEGATIVE	Here is a 1-2 sentence summary of the review: The Calm app has been a game-changer for my sleep, helping me fall asleep faster and more consistently, especially with the soothing sleep stories featuring Humphrey the cat. While the app can be a bit glitchy at times, the content and features have been a perfect solution for me.
a14aee39-cbe9-4d27-bf1c-3eb6c5e56182	Sunny L.	cancelled subscription, got charged next month.	1	0	6.67	########	NEGATIVE	0.9982309937477112	NEGATIVE	The user had a frustrating experience with the Calm app, as they cancelled their subscription but were still charged the following month.
f06a07d0-54db-4e04-b4e3-995013715237	D. Colbert	PURPOSELY FALSE & MISLEADING ADVERTISING: FAILED TO CLARIFY "SUBSCRIPTION REQUIRED" to hear anything they offer! ONCE AGAIN: NOT INTERESTED! Boycotting any apps attempting to force people to subscribe without a FREE TRIAL PERIOD to see if app even satisfies ones need(s). Your attempt is fuel for a TORT based LAWSUIT! And don't even think someone won't do it!	1	1		########	NEGATIVE	0.9997567534446716	NEGATIVE	Here is a 1-2 sentence summary of the review: The user was frustrated with the Calm app's misleading advertising, which failed to clearly state that a subscription is required to access its content. As a result, they felt deceived and are now boycotting the app, threatening to take legal action if similar tactics are used in the future.
a4e6f76e-d338-4917-a9c6-2021d88edfca	Rob Fann	The only way one can use this is through a paid subscription.	1	0	6.68.1	########	NEGATIVE	0.9995850920677184	NEGATIVE	Here is a summary of the review in 1-2 short sentences: The reviewer was disappointed to find that the Calm app requires a paid subscription to use, with no option for a free trial or basic version.
75c61fc0-0cf2-4b9c-8783-d7a045f8a594	Ann	waiting to try it tonite	4	0	6.44.3	########	NEGATIVE	0.9956243634223938	NEGATIVE	The user is looking forward to trying the Calm app tonight, indicating a sense of anticipation and eagerness to experience its benefits.
d735ee61-2b83-4e06-8c44-b647efda81a6	Louie Curl	it really helps me get to sleep and stay asleep	5	0	6.68.1	########	NEGATIVE	0.578149676	Neutral	Here is a summary of the review in 1-2 short sentences: The Calm app has been effective in helping the user fall asleep and stay asleep, providing a sense of relief and improved sleep quality.
										Here is a summary of the review:

Figure 2: CSV with positive or negative sentiment prediction along with professional summary of the user review

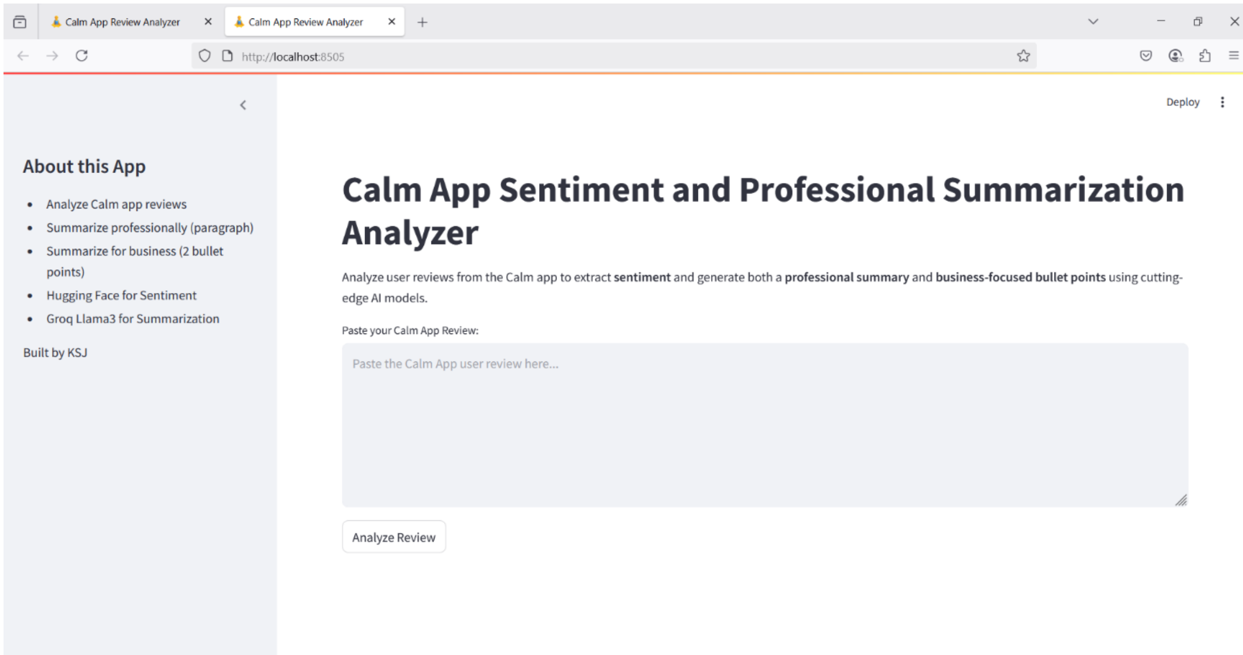


Figure 3: Streamlit app

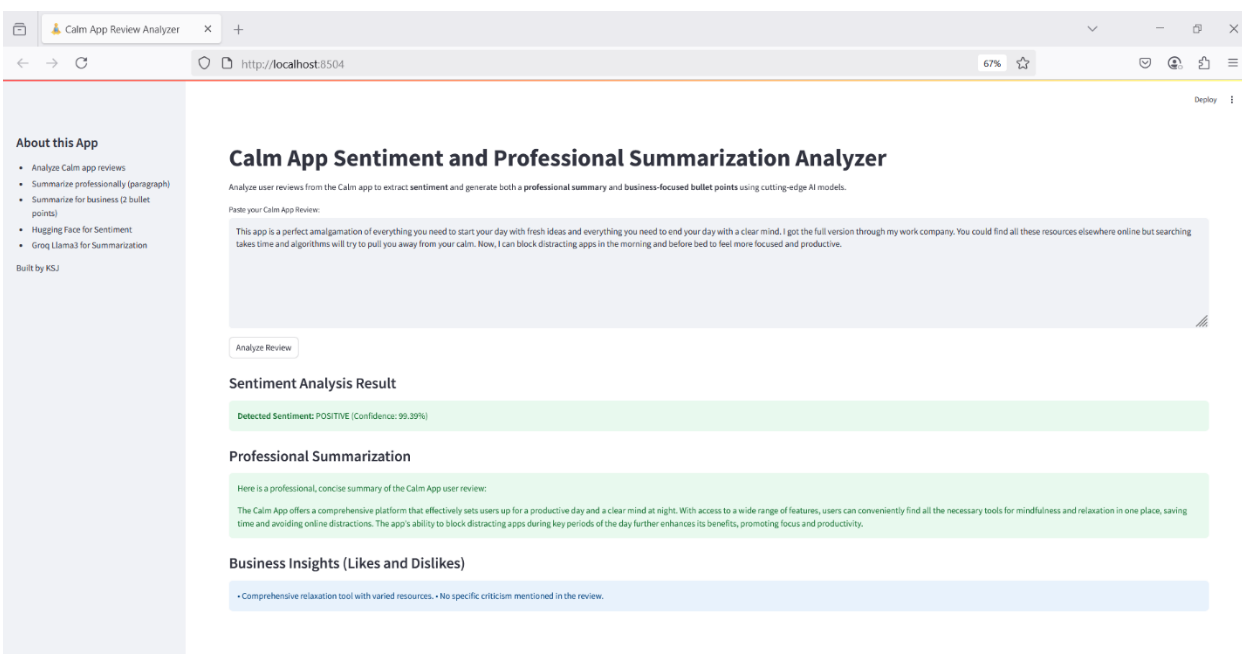


Figure 4: Results of the Positive Label and Business Insights Prediction, Summarization from Streamlit app

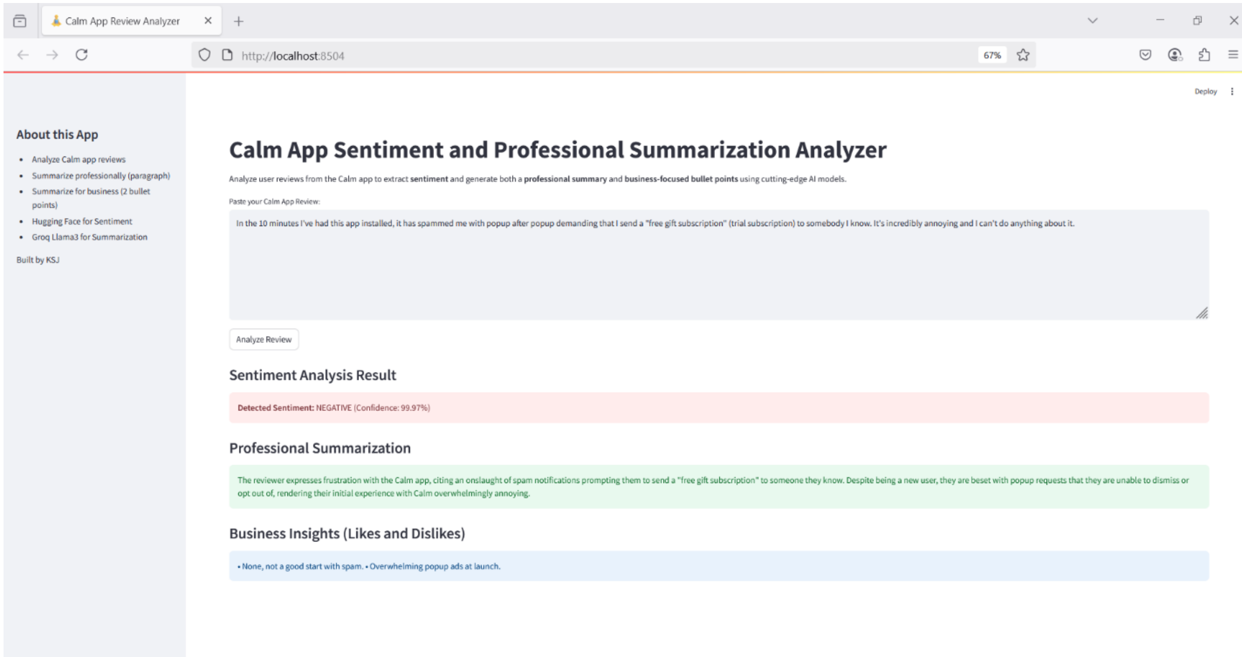


Figure 5: Results of the Negative Label and Business Insights Prediction, Summarization from Streamlit app

References:

ChatGPT- Writing and Paraphrasing