

Article

Efficient Sentiment, Emotion, and Toxic Speech Analysis on Social Media Using Lightweight LLMs: A Practical Approach for Digital Transformation

Ioannis Kapantaidakis *, Ioannis Kopanakis, Emmanouil Perakakis  and Konstantinos Vassakis 

Department of Management Science and Technology, Hellenic Mediterranean University, 72100 Agios Nikolaos, Greece; kopanakis@hmu.gr (I.K.); mperakakis@hmu.gr (E.P.); kvassakis@hmu.gr (K.V.)

* Correspondence: jkapad@hmu.gr

Abstract

Social media generates massive amounts of unstructured, multilingual textual data that contain many sentiments, emotions, and opinions from users that provide insights for enterprises undergoing digital transformation. In this paper, we evaluate a variety of commercially available, lightweight large language models (LLMs) for sentiment, emotion, aspect-based sentiment, and toxic speech detection on noisy, real-world social media datasets. These lightweight models are compared (against traditional RoBERTa-based classifiers and heavyweight “pro” LLMs) in terms of performance, latency, and cost trade-offs that are key for scalable deployment across the enterprise. Our results show that lightweight LLMs achieve good accuracy with considerably lower response times and costs, which makes them suitable as the next step in digital transformation for real-time social media analytics. Additionally, we investigate the importance of prompt engineering and show that preprocessing can be very limited for LLMs. This study provides operational guidelines regarding model selection based on latency, accuracy, and cost trade-offs as well as optimal prompt engineering techniques. Finally, it evaluates the intersection of operational performance and resource efficiency to help researchers and developers integrate LLM-based social media analytics into digitized business practices.

Keywords: large language models; social media analytics; sentiment analysis; digital transformation; prompt engineering; natural language processing



Academic Editor: María Dolores Ruiz

Received: 2 June 2026

Revised: 26 June 2026

Accepted: 6 July 2026

Published: 8 July 2026

Copyright: © 2026 by the authors.

Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. Licensee MDPI, Basel, Switzerland. This article

is an open access article distributed under the terms and conditions of the

[Creative Commons Attribution \(CC BY\)](#) license.

1. Introduction

Nowadays, in the digital era, corporations are increasingly leveraging information based on social media as part of the digital transformation of their organization. Social media express evolving groups of people’s moods, sentiments, and attitudes of customers, stakeholders, and the general public. This rich textual data can be used by enterprises to refine brand management, predict the market, enhance customer experience, speed up responses to new crises, and respond more quickly to emerging situations, all of which directly impact operational excellence and competitive advantage.

Sentiment analysis has evolved significantly over the last two decades. Early lexicon-based approaches were inexpensive and simple but struggled with understanding context, negation, and sarcasm [1]. Subsequent machine learning approaches required extensive quantities of labeled data and significant feature engineering [2]. More advanced patterns were detected by deep learning techniques that involved neural nets and embedded word

representations. Recently, transformers such as BERT and RoBERTa have applied attention to make context even more meaningful [3]. However, they require large computing power and huge amounts of well-labeled data, particularly for domain-specific tasks [4].

The arrival of large language models represents the largest recent development in sentiment and emotion analytics. Models such as GPT, Gemini, and Claude are capable of understanding and generating human-like language [5]. The most crucial property of these models is that they can use zero or a few examples to train [6]. This resolves the need for large-scale, expensive, high-quality labeled datasets required to train language models. This challenge is particularly evident in low-resource languages or domains with specific knowledge gaps [2]. LLMs can also be used to process social media text to identify sarcasm, slang, and mixed emotions that traditional models could not detect [7]. Yet, they require computing power (and thus substantial financial costs), there are potential biases, and detailed prompts to provide reliable outcomes [8]. These models can enable enterprises to perform sentiment, emotion, and toxic speech detection more accurately, faster, and more affordably. Balancing model performance with latency and financial costs is a key challenge for organizations undergoing digital transformation.

This research specifically aims to investigate the effectiveness of several popular commercial lightweight LLMs for performing sentiment and emotion-based analysis on social media data. Sentiment analysis, and more importantly, fine-grained emotion extraction, have been enormously valuable in today's social media environment. It enables companies, institutions, governments, and researchers to understand public opinions, brand reputation, emerging crises, and market trends. Simple sentiment classification can give a broad overview, but detecting emotion reveals more nuanced, complex, and sometimes contradictory emotional responses. Between a rapidly changing public opinion landscape on social media, the capability to dynamically, inexpensively track these transformations has become vital. However, social media text is often noisy and informal, and it is usually written in multiple languages, while in real-world scenarios, social media sentiment analysis should be performed multiple times (e.g., at regular intervals or in response to evolving events) to capture dynamic shifts in public opinion. Given all the above challenges, we aim to assess models that can achieve performance, accuracy, latency, and cost-effectiveness on these. We wish to help understand how these accessible LLMs work on real-world social media, especially compared with more traditional models.

This work conducts an examination of popular LLMs for performing sentiment and emotion analysis on noisy real-world multilingual social media data. We focus on models that are commonly accessible to researchers, small businesses, startups, as well as large enterprises seeking to optimize performance while reducing operational costs. By this, we help provide practical insights into when and how these affordable, low-latency models may become an alternative to much heavier and costly "pro" models while keeping performance competitive. Specifically, we present:

- An overview of how sentiment analysis methods have changed, focusing on the transition from older lexicon-based and machine learning techniques to deep learning and large language models;
- A comparison of popular commercial, lightweight LLMs on various public social media datasets, including performance metrics and response times;
- A comparison of LLM performance against traditional BERT-based model classifiers trained on complete data;
- A further study of how LLM performance is being affected by different prompting techniques;
- An examination if data preprocessing remains necessary with LLMs;

- A comparison of the selected lightweight LLMs performance, response time, and cost trade-offs against heavier, more expensive “pro” models.

2. Background and Related Work

2.1. Traditional Approaches to Sentiment Analysis

Sentiment analysis has been evolving through a number of steps before the arrival of large language models. In the early days, lexicon-based approaches were based on fixed sentiment lexicons, where words were assigned a polarity score from -5 to $+5$ [1,9–14]. SentiWordNet, AFINN, LIWC, and others used these dictionaries to measure text sentiment by counting positive and negative words. While simple and cost-effective, lexicon-based approaches were not effective in dealing with the language complexity of context, negation, sarcasm, and informal expressions found in social media.

To overcome these limitations, machine learning classifiers such as Support Vector Machines (SVMs), Naive Bayes, and Logistic Regression have been proposed to classify sentiment in n -grams and Term Frequency-Inverse Document Frequency (TF-IDF) [2,15,16]. These models, trained on annotated data, were able to identify sentiment patterns more reliably than lexicon-based methods and generalized better to unseen inputs. However, they still needed large labeled datasets and feature engineering and lacked deeper semantic understanding.

Deep learning helped in sentiment analysis as it learned hierarchical representations from raw text without manual feature design. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Gated Recurrent Units (GRU) enabled sequence modeling with better memory, overcoming earlier vanishing gradient problems [17–20]. Word embeddings like Word2Vec and GloVe also made semantic tagging much easier by placing words that are semantically related to each other well in vector space [19,20]. Finally, CNNs (convolutional neural networks) from image processing applications were trained to recognize the local features and phrases of words and thus provided valuable information for short text such as social media posts [21]. However, challenges remain in the form of longer dependencies and context understanding.

Transformer-based models achieved a breakthrough by processing whole text sequences simultaneously using self-attention mechanisms so that distance relationships between words could be better captured than in earlier models [3,22,23]. BERT and RoBERTa introduced bidirectional encoding, which enhanced the understanding of complex and ambiguous sentiments [23,24]. Extensions like DistilBERT and ALBERT made computation more efficient but still maintained performance. Despite this, transformers were expensive and labeled dataset-dominated, which was an issue for many organizations [4,25].

Together, these traditional stages—from lexicon-based to transformer models—laid the groundwork for the current state-of-the-art in sentiment analysis and paved the way for the powerful large language models that now enable more versatile and generalized natural language understanding.

2.2. Large Language Models (LLMs)

The introduction of large language models (LLMs) marks the latest and most significant advancement in sentiment and emotion analysis. Models such as GPT-3, GPT-4, Gemini, and Claude leverage transformer architectures trained on vast, diverse datasets, enabling zero-shot and few-shot learning without the need for extensive task-specific fine-tuning [5,6,26,27]. This flexibility allows LLMs to perform a wide range of downstream tasks—including sentiment classification, summarization, translation, and emotion detection—simply by modifying the input prompts [5].

Unlike traditional models, LLMs possess a superior ability to interpret complex language phenomena such as sarcasm, slang, and mixed emotions, which are prevalent in social media contexts but challenging for earlier approaches [7]. The advent of accessible open-source LLMs like LLaMA, Mistral, and Falcon has expanded experimentation opportunities beyond well-funded organizations, though large-scale commercial “pro” models remain resource-intensive [28]. However, operating these models entails significant computational costs, potential output ambiguity without precise prompting, risks of bias, and limited explainability [8,29,30].

Recent literature emphasizes the transformative impact of LLMs on sentiment analysis. Studies have demonstrated that zero-shot and few-shot LLMs can achieve competitive or superior results on certain classification tasks compared to traditional supervised models [6,31]. However, LLMs also exhibit weaknesses, particularly in fine-grained tasks such as aspect-based sentiment analysis and multi-emotion classification, where domain-specific models may sometimes outperform them [27,31]. The role of prompt engineering—crafting detailed and contextually rich instructions to guide model behavior—has emerged as a critical factor influencing LLM performance [32].

While comparative benchmarking of LLMs for sentiment analysis is well established in the literature, our work fills the gap in the literature by shifting the focus from theoretical maximum performance to practical operational viability. Our three major contributions are:

- **Focus on State-of-the-Art Lightweight Architectures:** Previous studies focus on heavier and more expensive LLMs (GPT-3.5, early LLaMA) or traditional transformers. We consider highly recent, cost-effective models (e.g., Gemini 2.0 Flash, Grok 3 mini, DeepSeek V3) that are very relevant to current developer ecosystems;
- **Comprehensive Multilingual and Multi-Task Scope:** Instead of isolating a task, our framework tests general sentiment, entity-specific aspect sentiment, fine-grained emotion, and toxic speech on manually annotated, multilingual social media datasets;
- **Operational Digital Transformation:** Unlike the current literature that ranks models according to classification accuracy, we map performance metrics against API latency and token costs. This results in a novel, practical way for businesses to identify the exact cost-performance trade-offs for real-time digital transformation.

3. Materials and Methods

In this section, we present the models, tasks, datasets, and evaluation measures that we researched. We categorize our evaluation tasks into four types: sentiment analysis, aspect-based sentiment analysis, toxic speech detection, and emotion analysis.

3.1. Overall Objective and Sub-Objectives

The overall objective of this study is to evaluate commercially available lightweight LLMs for real-time, multilingual social media analytics, focusing on their performance, latency, and operational cost to support digital transformation efforts in enterprises. To achieve this objective, the research addresses the following sub-objectives:

- **Performance Comparison:** Assess the accuracy, precision, recall, and F1-score of popular lightweight LLMs on sentiment analysis (SA), aspect-based sentiment analysis (ABSA), emotion analysis (EA), and toxic speech detection (TSD) tasks using diverse social media datasets;
- **Benchmarking:** Compare lightweight LLMs against traditional RoBERTa-based classifiers and heavyweight “pro” LLMs in terms of predictive quality, response time (latency), and computational cost;

- **Prompt Engineering:** Investigate the impact of different prompt engineering techniques, including Role-Playing (RP), Naming the Assistant (NA), Chain-of-Thought (COT), and Emotional Prompting (EP), on the models' classification performance;
- **Preprocessing Analysis:** Examine the necessity and influence of text preprocessing on LLM performance, particularly regarding handling noisy social media text;
- **Operational Trade-offs:** Analyze the interplay between model accuracy, latency, and cost to provide practical guidelines for scalable deployment in enterprise monitoring systems.

3.2. Sentiment Analysis

3.2.1. Sentiment Analysis Datasets

Sentiment analysis (SA) is a computational method used to determine the sentiment expressed in text. It typically analyzes text as positive, negative, or neutral. In our evaluation, we used the following datasets:

- **English Sentiment Analysis Dataset (ESAD):** A dataset of 1000 English tweets was created, half retrieved from Go et al. [33] and the other half from Yasserh [34];
- **Multilingual Sentiment Analysis Dataset (MSAD):** A subset of 1000 tweets was extracted from the multilingual dataset created by Barbieri et al. [35], with each tweet manually annotated as positive, negative, or neutral sentiment;
- **Manually Labeled Multilingual Sentiment Analysis Dataset (ML-MSAD):** To achieve a comprehensive analysis and ensure the models can handle previously unseen data, we manually collected and annotated a multilingual dataset containing about 1000 social media posts in various languages, primarily Spanish (40%), Greek (30%), and English (15%).

3.2.2. Aspect-Based Sentiment Analysis Datasets

Aspect-based sentiment analysis (ABSA) is a more detailed way to figure out what people think about specific things or aspects within a text. It goes beyond just saying whether a whole review is positive or negative. For example, if someone reviews a smartphone, a regular sentiment analysis might say the whole review is good. But ABSA could tell you that the camera got negative feedback, while the battery life got positive feedback. For our evaluation, the following dataset was used:

- **Aspect-Based Sentiment Analysis Dataset (ABSAD):** This is an entity-level sentiment analysis dataset of Twitter. The dataset includes three classes: positive, negative, and neutral [36].

3.3. Emotion Analysis

Emotion analysis (EA) represents a more granular progression of standard sentiment analysis. Whereas sentiment analysis attempts to classify text into large polarities such as positive, negative, or neutral, emotion analysis is an attempt to extract and classify more types of human emotions from texts. It enables us to reveal the total state of the person writing the emotion, helping us dig deeper into them and what they were really feeling and thinking. For example, instead of just labeling a text as negative, emotion analysis can determine whether an emotional response from that text is anger, sadness, or fear. The datasets we used for the evaluation are the following:

- **Emotion Analysis Dataset (EAD):** A collection of English tweets that have been annotated with six basic emotions: anger, fear, joy, love, sadness, and surprise [37];
- **Manually Labeled Emotion Analysis Multilingual Dataset (ML-EAD):** Similar to our sentiment analysis evaluation, we manually gathered and labeled 1000 social media

posts to identify four different emotions (joy, sadness, anger, fear). And a majority of these posts (40%) were in Spanish, followed by Greek (35%), followed by English (20%).

3.4. Toxic Speech Detection

Toxic speech detection (TSD) is a subset of natural language processing and machine learning that recently received substantial attention in the European Union and worldwide. Toxic speech detection refers to the process of automatically identifying and classifying online text that presents harmful, abusive, or offensive language targeting individuals or groups. For this task, we utilized the following dataset:

- YouTube Toxic comments: This dataset contains 1000 manually annotated toxicity comments in English, which were collected from YouTube videos about the 2014 Ferguson unrest [38].

3.5. Evaluation Models

For our study, we selected LLMs that are capable of good performance but still deliver a good trade-off between speed, low latency, and cost-effectiveness. This choice is motivated by the needs of real-world applications where models need to be fast, effective, and inexpensive. Even though academic research sets as a priority the theoretical perfection, there is a need for a useful and practical tool for real-world cases. According to this, our model selection includes:

- GPT-4o-mini;
- Gemini 2.0 Flash;
- DeepSeek V3;
- Grok 3 mini;
- Claude 3.5 Haiku.

Our comparison also extends to traditional BERT-based classifiers, for which we utilized the following RoBERTa-based models:

- twitter-XLM-roBERTa-base for sentiment analysis (SA-XLM): This is a multilingual roBERTa-based model trained on ~198 M tweets and finetuned for sentiment analysis over several languages;
- xlm-emo-t for emotion detection (EM-XLM): a multilingual RoBERTa-based model specifically trained to detect the following four emotions.

Both models are based on the XLM-T model presented in [35]. Finally, we wanted to evaluate the performance of the selected lightweight LLMs against heavier, more expensive “pro models” to determine the practical gains and losses in terms of overall performance, cost, and response time. For this case, we used:

- Gemini 2.5 Pro;
- Gemini 3 Pro.

3.6. Evaluation Metrics and Operational Significance

In order to translate our NLP evaluation to enterprise deployment scenarios, we operationalize “digital transformation” by translating these standard machine learning metrics into organizational Key Performance Indicators (KPIs). For an enterprise using social media analytics, the viability of an LLM is determined by three core business needs:

- Accuracy and F1-score (Reliability of Business Intelligence): Accuracy is the proportion of all responses that turned out right, while the F1-score is a balanced average that includes missed hits but also false alarms. These measures show how effectively we are capturing true public opinion, which directly affects the quality of decisions made in marketing, customer service, and crisis management;

- Precision and recall (Risk Mitigation Coverage): Precision indicates the number of times the model is correct when it labels something (avoiding “false alarms”), while Recall measures how much of the total relevant data the model actually found (avoiding “missed signals”). These metrics help organizations decide whether to prioritize accuracy (e.g., not wrongly flagging a polite comment as toxic) or thoroughness (e.g., ensuring every hint of a growing PR crisis is caught);
- Latency and Cost (Real-Time Workflow Integration): For true digital transformation, analytics must be integrated into high-throughput, automated pipelines. Response times determine if a model can support real-time dashboards, while cost-per-token determines the financial feasibility of continuous, large-scale data consumption.

3.7. Statistical Analysis

To ensure the reliability of our findings and account for variance, we conducted post hoc statistical analyses on the model predictions. For the overall accuracy metric we calculated 95% confidence intervals using non-parametric bootstrapping with 1000 resamples of the test set. Furthermore, to determine if the performance differences between the top-performing models were statistically significant, we applied McNemar’s test, adjusting for multiple comparisons where appropriate. A significance threshold of $p < 0.05$ was utilized for all tests.

4. Results

4.1. Sentiment Analysis Task

Below, we outline the evaluation metrics for the sentiment analysis tasks on both the English and multilingual datasets. Additionally, for the ESAD, we preprocess the text to investigate the impact of preprocessing on the results. For text preprocessing, we removed URLs, usernames, domains, and RT annotations.

Prompt used: Analyze the sentiment of the social media post. Output only one word: Positive, Negative, or Neutral. Post Text: [Text]

With the ESAD, Grok 3-mini achieved the highest raw accuracy (69%) and recall (69%) (Table 1). However, statistical analysis reveals that its numerical advantage over GPT-4o-mini ($p = 0.20$) and DeepSeek V3 ($p = 0.20$) is not statistically significant. Therefore, these three models effectively tie for top performance on the unprocessed English dataset. Claude 3.5 Haiku and Gemini 2.0 Flash performed significantly worse in this specific scenario ($p < 0.05$).

Table 1. Evaluation metrics for the SA task on the ESAD with no preprocessing.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
GPT-4o-mini	67%	68%	67%	68%	[0.64–0.70]	0.20 (ns)
Gemini 2.0 Flash	66%	67%	66%	66%	[0.63–0.69]	0.02
DeepSeek V3	67%	63%	67%	64%	[0.64–0.70]	0.20 (ns)
Grok 3-mini	69%	67%	69%	67%	[0.66–0.71]	Baseline
Claude 3.5 Haiku	52%	69%	52%	55%	[0.49–0.55]	0.00

Data preprocessing failed to meaningfully improve the top-performing models, with performance enhancements limited solely to Claude 3.5 Haiku, whose accuracy and F1-score increased marginally (Table 2). Conversely, preprocessing slightly degraded the performance of Grok 3-mini and GPT-4o-mini, while leaving the metrics for Gemini 2.0 Flash and DeepSeek V3 completely unaffected. In general, preprocessing minimized performance divergences but did not have a marked effect on model rankings, and Grok 3-mini maintained the highest accuracy (68%). This indicates that data preprocessing is largely unnecessary for highly capable LLMs processing noisy social media text.

Table 2. Evaluation metrics for the SA task on the ESAD with data preprocessing performed.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
GPT-4o-mini	66%	66%	66%	66%	[0.63–0.69]	0.27 (ns)
Gemini 2.0 Flash	66%	67%	66%	66%	[0.63–0.69]	0.18 (ns)
DeepSeek V3	67%	63%	67%	64%	[0.64–0.70]	0.71 (ns)
Grok 3-mini	68%	66%	68%	66%	[0.65–0.70]	Baseline
Claude 3.5 Haiku	55%	70%	55%	57%	[0.51–0.58]	0.00

According to the MSAD (Table 3), Grok 3-mini again achieved the highest absolute metrics (67% accuracy, 66% F1-score). However, the McNemar test confirms that GPT-4o-mini ($p = 0.28$), Gemini 2.0 Flash ($p = 0.45$), and DeepSeek V3 ($p = 0.28$) all performed within the statistical margin of error of the baseline. Claude 3.5 Haiku is less competent. Although the dataset is also more difficult in general, it shows slightly lower performance compared to other models.

Table 3. Evaluation metrics for the SA task on the MSAD.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
GPT-4o-mini	65%	65%	65%	65%	[0.62–0.68]	0.28 (ns)
Gemini 2.0 Flash	66%	65%	66%	65%	[0.63–0.68]	0.45 (ns)
DeepSeek V3	65%	64%	65%	64%	[0.62–0.68]	0.28 (ns)
Grok 3-mini	67%	67%	67%	66%	[0.64–0.70]	Baseline
Claude 3.5 Haiku	59%	65%	59%	59%	[0.56–0.62]	0.00

In the ML-MSAD (Table 4), we observe that the performance of all models drops further. The model accuracy ranged between 0.54 and 0.59, which means that the models were able to capture a little more than half of the sentiments that were observed in this dataset. F1-scores were similarly evenly distributed, suggesting a fair balance between precision and recall.

Table 4. Evaluation metrics for the SA task on the ML-MSAD.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
GPT-4o-mini	56%	56%	56%	56%	[0.54–0.60]	0.03
Gemini 2.0 Flash	55%	54%	55%	54%	[0.52–0.58]	0.00
DeepSeek V3	57%	58%	57%	58%	[0.54–0.60]	0.16 (ns)
Grok 3-mini	59%	59%	59%	59%	[0.56–0.62]	Baseline
Claude 3.5 Haiku	54%	58%	54%	54%	[0.51–0.57]	0.00
RoBERTa	56%	62%	56%	54%	[0.53–0.59]	0.06 (ns)

While Grok 3-mini performed best in raw (59% accuracy and F1-score), the performance of DeepSeek V3 ($p = 0.16$) and the traditional RoBERTa classifier ($p = 0.06$) are not statistically different from the Grok 3-mini baseline. This shows that for very noisy multilingual datasets, the domain-specific classifiers perform on par with the state-of-the-art lightweight LLMs, while the GPT-4o-mini, Gemini 2.0 Flash, and Claude 3.5 Haiku have a significant decrease in reliability ($p < 0.05$).

4.2. Aspect-Based Sentiment Analysis Task

The performance of the selected large language models on the ABSA task is shown in Table 5. Generally, performance is lower in ABSA than in SA since ABSA is a more complex and detailed task. General sentiment evaluation is used to categorize the sentiment from a text. By contrast, ABSA needs to locate specific aspects or entities within a text and then determine the sentiment expressed towards each of those individual aspects.

Table 5. Evaluation metrics for the ABSA task.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
GPT-4o-mini	61%	68%	61%	58%	[0.59–0.65]	0.16 (ns)
Gemini 2.0 Flash	62%	68%	62%	58%	[0.59–0.65]	0.17 (ns)
DeepSeek V3	61%	65%	61%	59%	[0.59–0.64]	0.13 (ns)
Grok 3-mini	62%	72%	62%	57%	[0.59–0.65]	0.39 (ns)
Claude 3.5 Haiku	63%	65%	63%	63%	[0.60–0.66]	Baseline

Prompt used: Analyze the sentiment of the social media post towards [keyword].
Output only one word: Positive, Negative, or Neutral. Post Text: [Text]

Claude 3.5 Haiku achieved the highest absolute accuracy (63%) and F1-score (63%), suggesting a highly balanced level of precision and recall. Gemini 2.0 Flash and GPT-4o-mini demonstrated highly comparable performance across all nominal metrics, both achieving an identical precision rate of 68% and matching F1-scores of 58%. However, applying statistical rigor fundamentally changes the interpretation of these results. Setting Claude 3.5 Haiku as the baseline reveals that the *p*-values for all other evaluated models exceed the 0.05 significance threshold. Consequently, there is no statistically significant performance difference among any of the evaluated models for the ABSA task.

4.3. Emotion Analysis Task

For this evaluation, we used two datasets: the English EAD and the manually annotated multilingual ML-EAD. The purpose was to identify specific emotions in social media posts. In EAD, six basic emotions were examined: sadness, joy, love, anger, fear, and surprise. The ML-EAD included four emotions: joy, sadness, anger, and fear. The models were instructed to display only the primary emotion, without additional explanations or references to other emotional states.

Prompt used for the EAD: Analyze the social media post and identify its emotion from the six categories only: sadness, joy, love, anger, fear, or surprise. Return only the primary emotion without any explanation or mention of secondary emotions. Post Text: [Text]

Prompt used for the ML-EAD: Analyze the social media post and identify its emotion from the four categories only: joy, sadness, anger, or fear. Return only the primary emotion without any explanation or mention of secondary emotions. Post Text: [Text]

Tables 6 and 7 show that the models achieve reduced F1 scores on the EA task compared to the sentiment analysis task, indicating that distinguishing a broader range of emotions is much more challenging. This also explains why models performed better on the ML-EAD (with four emotions) than on the EAD (which asked them to classify six different emotions).

Table 6. Evaluation metrics for the EA task on the EAD.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
GPT-4o-mini	58%	59%	58%	56%	[0.55–0.61]	Baseline
Gemini 2.0 Flash	56%	56%	56%	55%	[0.53–0.59]	0.17 (ns)
DeepSeek V3	56%	56%	56%	54%	[0.53–0.59]	0.14 (ns)
Grok 3-mini	56%	58%	56%	56%	[0.53–0.59]	0.20 (ns)
Claude 3.5 Haiku	54%	57%	54%	53%	[0.50–0.57]	0.00

Table 6 shows that the F1-scores for all models range from 0.53 to 0.56, suggesting that emotion analysis for six distinct categories is a significantly more complex task than binary or ternary sentiment classification. The closer range of scores also suggests a lesser performance distinction among models in comparison with the sentiment analysis tasks. GPT-4o-mini achieved the highest nominal metrics with an accuracy of 58% and an F1-

score of 56%. Yet, the statistical analysis indicates no significant difference between the performance of GPT-4o-mini and Grok 3-mini ($p = 0.20$), Gemini 2.0 Flash ($p = 0.17$), and DeepSeek V3 ($p = 0.14$). Only Claude 3.5 Haiku performed significantly worse on this specific task ($p = 0.00$). It seems that, in particular, it had the most difficulty classifying these six emotions. Like observations in other tasks, values of precision are generally slightly higher than recall for most models; in other words, they tend to take a more conservative approach with respect to positive predictions (fewer false positives) but may miss some (less recall).

Table 7. Evaluation metrics for the EA task on the ML-EAD.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	<i>p</i> -Value
GPT-4o-mini	74%	82%	74%	76%	[0.71–0.77]	0.00
Gemini 2.0 Flash	80%	82%	80%	80%	[0.78–0.83]	Baseline
DeepSeek V3	63%	72%	63%	59%	[0.60–0.66]	0.00
Grok 3-mini	76%	78%	76%	74%	[0.73–0.78]	0.00
Claude 3.5 Haiku	68%	75%	68%	65%	[0.64–0.71]	0.00
RoBERTa	62%	63%	62%	61%	[0.59–0.65]	0.00

We observe in Table 7 that for most LLMs, the performance of the ML-EAD is much better than that of the EAD, possibly due to the smaller classes that are available for emotion classification. Gemini 2.0 Flash also outperformed with the best accuracy (0.80) and F1-score (0.80), demonstrating strong performance for multilingual emotion detection with fewer categories. This victory is statistically significant across the board since every other model generated a p -value of 0.00 against the Gemini baseline. GPT-4o-mini had a very high F1-score (0.76) as well. Grok 3-mini performed competitively with an F1-score of 0.74. DeepSeek V3 and Claude 3.5 Haiku's F1-scores of 0.59 and 0.65 were obtained, respectively, below the other LLMs.

Lastly, RoBERTa also achieved reasonably good performance on this multilingual setup with an F1-score of 0.61. When the task was converted to four emotions in multilingual language, Gemini 2.0 Flash and GPT-4o-mini performed especially well, which can be blamed on better efficiency based on having reduced the number of emotion categories.

4.4. Toxic Speech Detection Task

For the toxic speech detection task, we evaluated the performance of the selected LLMs using the YouTube Toxic Comments Analysis Dataset, and the results are presented in Table 8. This task aims to identify and classify online textual content containing harmful, abusive, or offensive language.

Table 8. Evaluation metrics for the TSD task.

Model	Accuracy	Precision	Recall	F1-Score	95% CI	<i>p</i> -Value
GPT-4o-mini	84%	84%	84%	84%	[0.82–0.87]	Baseline
Gemini 2.0 Flash	82%	83%	82%	82%	[0.80–0.84]	0.03
DeepSeek V3	82%	82%	82%	82%	[0.80–0.84]	0.04
Grok 3-mini	84%	84%	84%	84%	[0.82–0.86]	0.69 (ns)
Claude 3.5 Haiku	77%	80%	77%	76%	[0.74–0.79]	0.00

Prompt used: Classify if the following YouTube comment is toxic. Respond with only TRUE if toxic, FALSE if not. Comment Text: [Text]

The models had decent performance in detecting toxic speech in general, possibly because it is a simple task with just two labels to label. GPT-4o-mini and Grok 3-mini both achieved the highest raw metrics, each securing an 84% accuracy, 84% precision,

84% recall, and an 84% F1-score. The statistical analysis ($p = 0.69$) confirms there is no meaningful difference between these two top-tier models, establishing a tie for first place. DeepSeek V3 and Gemini 2.0 Flash both came into the mid-range with similar performance, with F1-scores of 0.82 and 0.82, respectively. Both models achieved a good balance of all measures. In comparison, Claude 3.5 Haiku achieved the lowest performance of the tested LLMs with an accuracy of 0.77 and an F1-score of 0.76, suggesting it was found to perform worse on this task when compared with the rest. The results showed that although several LLMs are capable of detecting toxic speech, the level of effectiveness is decidedly different between these LLMs.

4.5. Prompt Engineering

In this section, we aim to investigate whether LLMs' performance is affected by different prompting techniques. Similar to previous research [32], we used the following prompt engineering techniques on the ESAD:

1. **Chain-of-Thought (COT):** This technique guides the model to follow a step-by-step reasoning process for evaluating the response.

Prompt used: Analyze the sentiment of the following social media post. To ensure accuracy, follow this step-by-step reasoning process: First, identify key phrases and words in the post that carry emotional weight (e.g., "amazing" for positive, "awful" for negative). Second, assess the context, including emojis, hashtags, or slang that may amplify or alter the tone (e.g., sarcasm or irony). Finally, balance the positive and negative elements to classify the sentiment as positive, negative, or neutral. Social media post: [text]

2. **Emotional Prompting (EP):** This method appends a psychological phrase to the prompt, acting as an emotional stimulus for the LLM.

Prompt used: Classify the sentiment of this social media comment as positive, negative, or neutral, focusing on the language and tone used. Your thoughtful analysis could make a huge difference in understanding online trends, and it should be as accurate as possible. Social media post: [text]

3. **Role-Playing (RP):** This strategy provides a persona's description in the prompt, encouraging the LLM to adopt a specific character and perspective for the task. Prompt used: You are Dr. Langston, a leading social media analytics expert with a PhD in computational linguistics and a passion for decoding digital emotions. Your work helps brands understand their audience's feelings with precision and empathy. In your role, analyze the sentiment of the following social media post, classifying it as positive, negative, or neutral. Social media post: [text]

4. **Naming the Assistant (NA):** This involves giving the LLM a nickname, often used in conjunction with Role-Playing by adding the name to the beginning of the instruction.

Prompt used: You are SIA, an AI that understands social media emotions, using NLP expertise and knowledge of human expression. As SIA, evaluate the sentiment of this social media post to determine if it is positive, negative, or neutral. Social media post: [text]

GPT-4o-mini performs best with NA prompting, while the accuracy and F1-score for COT and EP show a significant drop (Table 9). Precision is the highest with COT (0.71).

Table 9. Evaluation metrics across prompting techniques for GPT-4o-mini.

Prompt	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
COT	60%	71%	60%	62%	[0.57–0.63]	0.00
EP	60%	69%	60%	62%	[0.57–0.63]	0.00
RP	62%	68%	62%	64%	[0.59–0.66]	0.00
NA	66%	70%	66%	67%	[0.63–0.69]	Baseline

Gemini 2.0 Flash performs consistently across prompts, but NA is a bit better than the others. Precision was best in COT (0.68). Grok 3-mini performs very well with RP and NA, indicating good robustness across prompting techniques. The precision of all prompts remains unchanged (0.67), which shows stable, good prediction quality (Tables 10 and 11). DeepSeek V3 shines in RP (Table 12) but struggles significantly with EP, experiencing a notable decrease in performance. However, it is worth noting that GPT-4o-mini actually experienced the largest overall performance drop when utilizing EP. Precision is unchanged (0.68) among all prompts. Claude 3.5 Haiku performs poorly across all prompts (Table 13), with COT being particularly ineffective (accuracy 0.45). RP produced the strongest outcomes; however, compared to other models, the metrics are still low.

Table 10. Evaluation metrics across prompting techniques for Gemini 2.0 Flash.

Prompt	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
COT	62%	68%	62%	64%	[0.60–0.65]	0.01
EP	63%	66%	63%	64%	[0.60–0.66]	0.02
RP	64%	65%	64%	64%	[0.62–0.67]	0.50 (ns)
NA	65%	66%	65%	65%	[0.62–0.68]	Baseline

Table 11. Evaluation metrics across prompting techniques for Grok 3 mini.

Prompt	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
COT	68%	67%	68%	67%	[0.65–0.71]	0.35 (ns)
EP	67%	67%	67%	67%	[0.64–0.70]	0.02
RP	69%	67%	69%	68%	[0.66–0.72]	Baseline
NA	69%	67%	69%	68%	[0.66–0.72]	0.79 (ns)

Table 12. Evaluation metrics across prompting techniques for DeepSeek V3.

Prompt	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
COT	65%	68%	65%	66%	[0.62–0.68]	0.55 (ns)
EP	61%	68%	61%	63%	[0.59–0.64]	0.00
RP	66%	68%	66%	67%	[0.63–0.69]	Baseline
NA	65%	68%	65%	66%	[0.62–0.68]	0.57 (ns)

Table 13. Evaluation metrics across prompting techniques for Claude 3.5 Haiku.

Prompt	Accuracy	Precision	Recall	F1-Score	95% CI	p-Value
COT	45%	68%	45%	47%	[0.42–0.48]	0.00
EP	51%	69%	51%	53%	[0.48–0.54]	0.00
RP	55%	67%	55%	57%	[0.52–0.57]	0.53 (ns)
NA	54%	68%	54%	56%	[0.51–0.57]	Baseline

RP and NA were, overall, the best types of prompts among models, as RP provided a boost to Grok 3-mini, DeepSeek V3, and Claude 3.5 Haiku, and NA showed great improvement for GPT-4o-mini and Gemini 2.0 Flash (Figure 1). COT appears to be the least effective for most models, especially Claude 3.5 Haiku, indicating that Chain-of-Thought prompting may not be an appropriate model for the task. EP has mixed results, performing poorly on DeepSeek V3 and Claude and moderately on others.

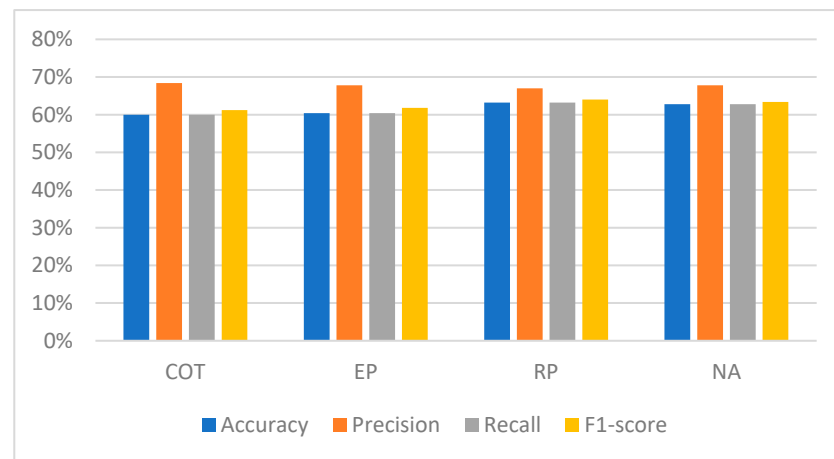


Figure 1. Average metrics across models by prompting technique.

RP provided the highest average accuracy (0.63), recall (0.63), and F1-score (0.64), which means that it is the most effective prompting method. NA demonstrates surprisingly strong performance, achieving metrics very close to RP (F1-score 0.64, for example). In contrast, COT has the lowest average accuracy (0.6), recall (0.6), and F1-score (0.612), although it has higher precision (0.68). This suggests that COT may produce fewer false positives than other techniques but can fall short of general correctness and coverage. EP is a bit better than COT, but remains inferior to RP and NA with balanced metrics (~0.60–0.68). RP is the most effective prompting technique, followed by NA.

To address the task-dependency of these instructional strategies, we also applied the prompt evaluation to the Aspect-Based Sentiment Analysis (ABSA) dataset with GPT-4o-mini as the representative lightweight architecture. For this evaluation, we adapted the four prompting techniques to specifically target entity-level sentiment extraction:

1. Chain-of-Thought (COT):

Prompt used: Analyze the sentiment of the social media post towards the specific entity: [keyword]. To ensure accuracy, follow this step-by-step reasoning process: First, locate the exact mention of [keyword] in the text. Second, identify words or phrases directly describing or relating to [keyword] and evaluate their emotional weight. Finally, isolate these elements from the rest of the text to classify the sentiment specifically directed at [keyword] as Positive, Negative, or Neutral. Output only one word. Social media post: [text]

2. Emotional Prompting (EP):

Prompt used: Analyze the sentiment of this social media post towards the entity: [keyword]. Output only one word: Positive, Negative, or Neutral. Your thoughtful and precise analysis could make a huge difference in helping us understand specific online consumer trends, and it should be as accurate as possible. Social media post: [text]

3. Role-Playing (RP):

Prompt used: You are Dr. Langston, a leading social media analytics expert with a PhD in computational linguistics and a passion for decoding digital emotions. Your work helps brands understand their audience's feelings toward specific products and topics with precision and empathy. In your role, analyze the sentiment of the following social media post towards the entity: [keyword]. Output only one word: Positive, Negative, or Neutral. Social media post: [text]

4. Naming the Assistant (NA):

Prompt used: You are SIA, an AI that understands targeted social media emotions, using NLP expertise and knowledge of human expression. As SIA, evaluate the sentiment of this social media post towards the entity: [keyword] to determine if it is Positive, Negative, or Neutral. Output only one word. Social media post: [text]

The results were strikingly different from general sentiment classification (Table 14). For the complex reasoning task, COT was the most effective prompt with an F1-score of 61% (95% CI: 0.59–0.65). Emotional Prompting and Role-Playing performed very similarly to COT ($p = 0.76$ and $p = 0.21$, respectively), while Naming the Assistant (NA) performed much worse, falling to an F1-score of 55% ($p = 0.02$). This indicates a complete performance inversion compared to the ESAD results, where NA was optimal and COT struggled.

Table 14. Evaluation metrics across prompting techniques for GPT-4o-mini and ABSA task.

Prompt	Accuracy	Precision	Recall	F1-Score	95% CI	<i>p</i> -Value
COT	62%	64%	62%	61%	[0.59–0.65]	Baseline
EP	62%	68%	62%	59%	[0.59–0.65]	0.76 (ns)
RP	61%	67%	61%	57%	[0.58–0.64]	0.21 (ns)
NA	59%	67%	59%	55%	[0.56–0.62]	0.02

4.6. Response Time Analysis Insights

The following figure (Figure 2) displays the average response time taken for each evaluation task recorded in milliseconds (ms). This work provides an important benchmark for the operational efficiency and effectiveness of these commercially available LLMs for use in real-time social media analysis.

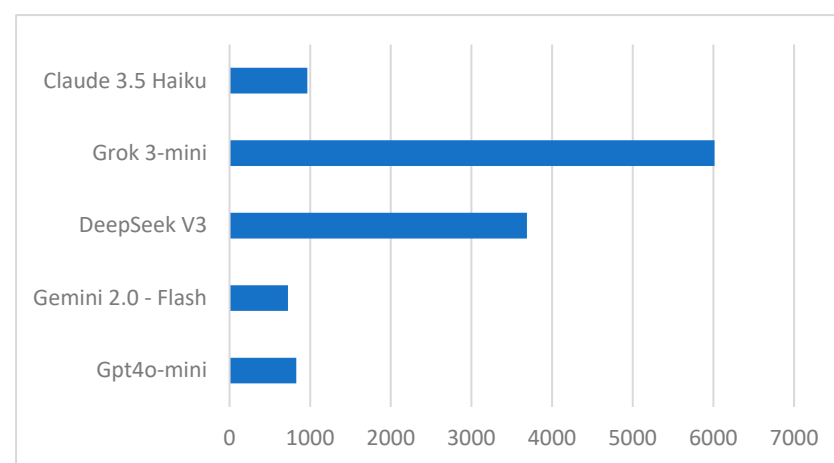


Figure 2. Average response time (ms) per model.

Gemini 2.0 Flash is the fastest model, averaging around 724.68 ms with response times close to GPT-4o-mini (827.80 ms), followed by Claude 3.5 Haiku (964.07 ms). These three models consistently respond in under one second, which makes them ideal for high-throughput, low-latency applications such as monitoring social media feeds. DeepSeek V3 and Grok 3-mini, respectively, have longer average response times. The average response time of DeepSeek V3 is approximately 3.7 s. Grok 3-mini is the slowest, with more than 6 s. Their longer response times suggest that, despite having highly competitive accuracy, their computation and structure make them weak when implemented in use cases requiring fast response times. These response times should be considered in conjunction with the performance as previously discussed. For example, Grok 3-mini frequently performed better in the sentiment tasks, yet its relatively high response time means it may be in a compromised speed vs. accuracy trade-off. On the other hand, in fact, Gemini 2.0 Flash

manages a nice balance between fairly good performance and very fast response times. In conclusion, for applications that require fast and efficient handling of social media data, Gemini 2.0 Flash, GPT-4o-mini, and Claude 3.5 Haiku are the most desirable options thanks to their fast return times. When choosing an LLM for social media analysis, it is important to find a suitable balance between the accuracy needed by the required algorithm and the acceptable latency and performance for the exact use case.

4.7. LLM Performance vs. Gemini Pro Models

Moreover, we compared the performance of our selected lightweight large language models to the heavier and more expensive Gemini Pro variants, evaluating performance trade-offs in accuracy, response time, and cost of computation. We measured these Pro models for sentiment analysis, aspect-based sentiment analysis, and emotion analysis tasks. Overall, Pro models performed well on different sentiment and emotion analysis tasks, outperforming their lighter competitors. However, their overall superiority was not as clear as one might have expected. Figure 3 compares the performance of the Pro models against the best lightweight LLM from each task.

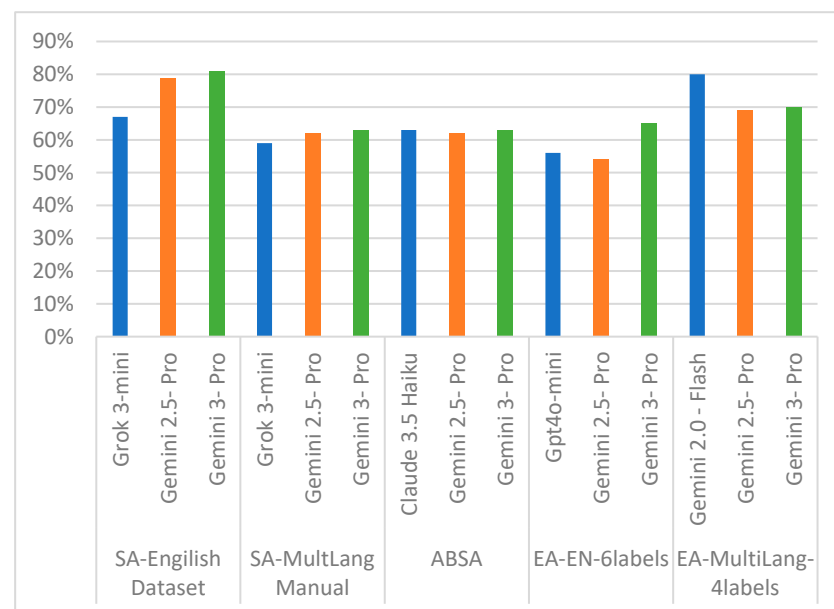


Figure 3. Pro model comparison vs. best lightweight LLM per task (F1-score performance).

On the ESAD, Gemini 3 Pro achieved the highest F1-score at 0.81, closely followed by Gemini 2.5 Pro at 0.79. Similarly, in the more complex 6-label EAD, Gemini 3 Pro compared favorably with all other models, all lightweight, and Gemini 2.5 Pro, attaining an F1-score of 0.65. Meanwhile, the pro models had also been getting good results in the ML-MSAD, with the Gemini 3 Pro having 0.63 F1 and the Gemini 2.5 Pro being at 0.62 F1-score, making them the best in this challenging dataset.

But the comparison also showed cases where the performance gap significantly narrowed, or even reversed, challenging the concept that pro models are always way better. For the ABSA task, we showed a very high F1-score of 0.63 among Gemini 3 Pro against Claude 3.5 Haiku, which is a lightweight model with good value for money, proving that a slightly more moderate model can achieve the same value. In addition, on the ML-EAD, the lightweight Gemini 2.0 Flash ranked first with an F1-score of 0.80 and outperformed Gemini 2.5 Pro (0.69 F1-score) and Gemini 3 Pro (0.70 F1-score). This indicates that for multilingual emotion detection with a constrained label set, highly optimized lightweight

models like Gemini 2.0 Flash not only match but actively outperform the capabilities of expensive ‘pro’ models.

Furthermore, our results point to an important trade-off. While “pro” models are typically able to achieve a higher accuracy with more capabilities, especially under challenging situations, they come at a higher computational cost—and, in turn, longer response times (Figure 4). For example, Gemini 2.5 Pro responded on average in 6375.25 ms, while Gemini 3 Pro contributed on average in 8328.5 ms, significantly behind the lightweight-response-time winner, Gemini 2.0 Flash, which averaged 724.68 ms.

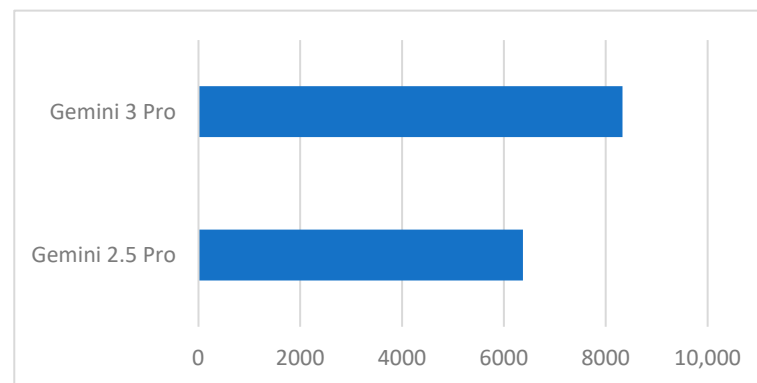


Figure 4. Average response time (ms) for “Pro” models.

4.8. Cost Analysis

Cost efficiency is an important key for enterprises in the field of large language models (LLMs) for highly scalable social media analytics that can be used to foster digital transformation and minimize costs.

Gemini Pro models come at a higher price than lightweight LLMs. Specifically, the expense of Gemini 2.5 Pro is \$1.25 for input and \$10 for output per 1 million tokens, while the cost of Gemini 3 Pro is \$2 for input and \$12 for output. As a result, Gemini 3 Pro and Gemini 2.5 Pro for typical requests (e.g., 1000 input and 500 output tokens) carry an external cost 25 and 20 times higher than Gemini 2.0 Flash, respectively. Since monitoring social media requires processing tons of data, these costs add up fast; therefore, such models could not be used by small/medium businesses. On the other hand, lightweight models including GPT-4o-mini, Gemini 2.0 Flash, and DeepSeek V3 offer substantially reduced costs, making it actually affordable for continuous real-time social media analytics.

Although “pro” models can deliver better accuracy in many cases, their high cost, slow response time, and economic constraints limit their practical application in business environments. Lightweight LLMs offer advantages, including reduced computational costs, making them suitable for organizations seeking to make advanced AI-driven social media analytics more accessible, thereby facilitating digital transformation while optimizing resource usage.

5. Discussion

In this study, we analyzed the performance of several popular commercial large language models on sentiment and emotion analysis on social media data and evaluated their capabilities in different tasks and prompting systems. In particular, our aim was to compare their performance against real-world tasks that depend on speed, latency, and cost.

5.1. LLM Performance Comparison and Response Times

A comprehensive comparison of the model performances is summarized in Table 15. While Grok 3-mini frequently achieved the highest accuracy and F1-scores in general

sentiment analysis tasks, our statistical analysis demonstrates that its performance is not significantly different from GPT-4o-mini, DeepSeek V3, and Gemini 2.0 Flash. Similarly, in the aspect-based sentiment analysis task, all evaluated models performed within the statistical margin of error of one another, despite Claude 3.5 Haiku posting the highest raw numbers. However, distinct performance gaps emerged in emotion analysis. While the complex 6-emotion classification task resulted in a multi-model tie, Gemini 2.0 Flash emerged as the undisputed, statistically significant champion for the constrained 4-emotion multilingual dataset (ML-EAD).

Table 15. Model performance comparison per task.

Evaluation Task	Statistically Significant Winner (s)	Top Raw Performer	Reasoning
Sentiment Analysis	Grok 3-mini, GPT-4o-mini, DeepSeek V3, Gemini 2.0 Flash	Grok 3-mini	Grok 3-mini frequently achieved the highest raw metrics across English and multilingual datasets, but statistically ties with the other top lightweight models.
Aspect-Based Sentiment Analysis	All Models	Claude 3.5 Haiku	Claude 3.5 Haiku achieved the highest raw accuracy and F1-score (63%), but all evaluated models performed within the statistical margin of error of one another.
Emotion Analysis	Gemini 2.0 Flash (4-class) Statistical Tie (6-class)	Gemini 2.0 Flash GPT-4o-mini	Gemini 2.0 Flash is the top performing model for 4-class multilingual datasets. The complex 6-class dataset resulted in a multi-model tie.
Toxic Speech Detection	GPT-4o-mini, Grok 3-mini	GPT-4o-mini, Grok 3-mini	Both models tied with identical raw metrics (84% F1). However, GPT-4o-mini is the definitive operational winner here due to its drastically lower latency.
Overall Response Time	Gemini 2.0 Flash	Gemini 2.0 Flash	Stands out as the quickest model with an average response time of ~724.68 ms, followed closely by GPT-4o-mini (~827.80 ms), making them ideal for real-time dashboards.

For response time efficiency, Gemini 2.0 Flash ranked as the quickest model, with an average response time of about 724.68 ms, followed by GPT-4o-mini and Claude 3.5 Haiku, which each delivered responses in less than one second. DeepSeek V3 and Grok 3-mini had significantly longer average response times, at about 3688.38 ms and 6014.41 ms, respectively. This points to a trade-off, where Grok 3-mini models may be better performing on certain tasks but have higher latency. As a result, model selection should align with the specific criteria of each application.

5.2. LLM Performance vs. Traditional RoBERTa-Based Models

Compared with legacy RoBERTa-based model classifiers, LLMs were competitive and, in most cases, performed well, particularly on challenging multilingual data with manual annotations. In the manually labeled multilingual sentiment analysis dataset, the traditional RoBERTa classifier did not simply fall into the “middle tier”; rather, it performed statistically on par with the top-performing lightweight LLMs (Grok 3-mini and DeepSeek V3). This represents that traditional models are still effective while LLMs provide more advanced capabilities, especially for multilingual environments that frequently struggle with finding labeled data.

5.3. Impact of Different Prompting Techniques

The choice of prompting strategy significantly influenced model performance depending on the complexity of the task. For straightforward, general sentiment classification (ESAD), persona-based approaches like Role-Playing and Naming the Assistant proved most effective for lightweight models. Conversely, Chain-of-Thought (COT) often degraded performance on these simple tasks. This underperformance aligns with recent literature suggesting that forcing explicit reasoning steps introduces verbosity that can interfere with constrained, single-label classification outputs [39]. However, when evaluating the more complex aspect-based sentiment analysis (ABSA) task using GPT-4o-mini, COT became the optimal strategy ($F1 = 61\%$), while Naming the Assistant resulted in a statistically significant performance drop ($F1 = 55\%$, $p = 0.02$). This performance inversion confirms the theoretical hypothesis that COT is primarily beneficial for multi-step reasoning tasks, such as locating a specific entity and isolating its associated sentiment, rather than basic classification. This critical finding demonstrates that there is no universal prompt template for social media analytics, highlighting the need for custom prompt engineering to enhance LLM performance in tailored scenarios.

5.4. Necessity of Data Preprocessing with LLMs

Preprocessing social media text improved mostly slight performance of the models, mainly for low-performance models like Claude 3.5 Haiku. For top-tier models like Grok 3-mini, GPT-4o-mini, Gemini 2.0 Flash, and DeepSeek V3, noise removal had either no effect or slightly degraded their metrics. This indicates that LLMs allow powerful language understanding capabilities and thus that preprocessing steps are minimally impactful, meaning they can process raw, noisy social media data without much cleaning.

5.5. Performance, Response Time, and Cost Trade-Offs: Lightweight vs. Pro LLMs

Our study compared lighter lightweight LLMs against heavier and more expensive pro models (Gemini 2.5 Pro and Gemini 3 Pro) on sentiment analysis, aspect-based sentiment analysis, and emotion analysis tasks. Pro models tended to perform reasonably well, frequently outperforming lightweights, but not as well as expected. Specifically, pro models outperformed lightweight LLMs in sentiment analysis, aspect-based sentiment analysis, and emotion analysis tasks. But the advantage of this was far from apparent. Gemini 3 Pro matched Claude 3.5 Haiku in aspect-based sentiment analysis, while lightweight Gemini 2.0 Flash performed better than pro models for multilingual emotion analysis with fewer categories. Pro models also showed far longer response times and cost approximately 5–25× that of lighter ones. This suggests a higher computing demand on their computational capability. While pro models have better overall accuracy, lightweight LLMs offer a compromise in terms of performance, efficiency and price point on the high end for complex social media workloads.

5.6. Enterprise Deployment

In this research, we illustrate that digital transformation via LLMs is not only about adopting the most powerful AI but also about selecting the architecture that best fits into specific organizational workflows. Based on our results, we propose the following deployment mapping:

- **Real-Time Monitoring Dashboards:** For continuous ingestion workflows (live brand-sentiment tracking), Gemini 2.0 Flash is an ideal engine. Due to its sub-second latency, the data can be visualized without bottlenecking API pipelines;
- **Automated Customer Support:** When companies automatically route toxic or highly negative comments to human representatives, GPT-4o-mini is optimal. While it statis-

tically tied with Grok 3-mini for toxic speech detection accuracy, its drastically lower latency (~827 ms versus Grok's ~6014 ms) ensures that customer service workflows are streamlined without introducing system delays;

- **High-Fidelity Sentiment Archiving:** For applications where accuracy is vital while latency is secondary (e.g., end-of-day sentiment aggregation), Grok 3-mini provides the most reliable general sentiment classification. Organizations can capitalize on its top-ranking nominal raw metrics (69% accuracy) for general sentiment classification, making its 6 s response time an acceptable trade-off;
- **Batch-processed market research:** When tracking specific product features (e.g., weekly product-aspect reports), Claude 3.5 Haiku is recommended. Although aspect-based sentiment extraction resulted in a statistical tie across all evaluated models, Claude 3.5 Haiku achieved the highest absolute balanced metrics (63% F1-score), providing the nuanced data required to inform long-term strategy planning.

By bringing different lightweight models into alignment with these different deployment scenarios, organizations can deliver on their digital transformation objectives in real time without losing computational efficiency but with actionable business intelligence.

6. Conclusions

In this study, we evaluated sentiment, emotion, aspect-based sentiment and toxic speech on real-world, multilingual datasets and showed that lightweight models provide a very strong, inexpensive replacement for resource-heavy “pro” variants and traditional classifiers. Crucially, we demonstrated that in specific scenarios, such as multilingual emotion detection, highly optimized lightweight models like Gemini 2.0 Flash can actually outperform vastly more expensive ‘pro’ models.

Our analysis highlights critical operational trade-offs among the evaluated models, showing that while several architectures achieve statistical parity in predictive accuracy, their execution metrics are quite different. Although Grok 3-mini had the highest nominal raw metrics for general sentiment analysis and Claude 3.5 Haiku led in raw aspect-based sentiment metrics, statistical analysis confirms that both tasks resulted in multi-model ties where these models did not significantly outperform their peers. Consequently, Gemini 2.0 Flash and GPT-4o-mini are the optimal all-round selections for high-throughput enterprise pipelines. This choice is driven by their highly competitive, top-tier statistical accuracy, Gemini's clear victory in constrained multilingual emotion tracking, and their sub-second execution time (compared with Grok 3-mini's six-second processing time).

Furthermore, our study into prompt engineering suggests that the approach is task-dependent rather than generalizable. Persona-based approaches like Role-Playing and Naming the Assistant optimize straightforward, single-label sentiment classification, but a complete performance inversion occurs for more complicated tasks such as aspect-based sentiment analysis, where Chain-of-Thought reasoning is statistically required to guide the model's isolating logic.

These results, from a managerial perspective, provide valuable insights for selecting and running AI-powered social media analytics tools that can help in monitoring public sentiment dynamically, increase the effectiveness of digital marketing efforts, and proactively mitigate reputational threat risks. Moreover, the demonstrated robustness of these models across multiple languages is beneficial for global businesses that require inclusive and market-sensitive insights.

In conclusion, lightweight LLMs are becoming the enablers of operational excellence and agile decision-making in digital transformation journeys. With businesses increasingly turning to social data to guide their strategy and operations, investing in effective, accurate,

and cost-effective AI analytics solutions will be key to driving competitive advantage and innovation in the digital age.

While this study presented that lightweight LLMs are effective in social media analytics, there are still some directions for future research. First, future work should explore the integration of multimodal data. The current study has only been based on text data, but social media posts often contain text with images, audio, and videos. Second, subsequent research could expand the multilingual evaluation. The present manually labeled datasets primarily consist of English, Greek, and Spanish texts. Extending this to more diverse, low-resource languages is further proof of the global applicability of these models. In addition, while we were able to assess latency and cost in a structured experimental setup, testable lightweight models in real-time enterprise dashboards would provide a good case study of their performance to show that these lightweight models can be successfully implemented in real-time operations.

Author Contributions: Conceptualization, I.K. (Ioannis Kapantaidakis), I.K. (Ioannis Kopanakis) and E.P.; methodology, I.K. (Ioannis Kapantaidakis) and I.K. (Ioannis Kopanakis); software, I.K. (Ioannis Kapantaidakis) and E.P.; validation, E.P., K.V. and I.K. (Ioannis Kapantaidakis); data curation, I.K. (Ioannis Kapantaidakis); writing—original draft preparation, E.P. and I.K. (Ioannis Kapantaidakis); writing—review and editing, K.V. and I.K. (Ioannis Kapantaidakis); visualization, E.P.; supervision, I.K. (Ioannis Kopanakis). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The public datasets analyzed in this study are openly available in third-party repositories. The Twitter Tweets Sentiment Dataset (analyzed within ESAD) is available at <https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset> (accessed on 5 March 2026). The Twitter Entity Sentiment Analysis dataset (ABSAD) is available at <https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis> (accessed on 5 March 2026). The Emotions dataset (EAD) is available at <https://www.kaggle.com/datasets/nelgiriyeewithana/emotions> (accessed on 5 March 2026). The YouTube Toxicity Data is available at <https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data> (accessed on 5 March 2026). The custom, newly annotated multilingual datasets generated specifically during this study—namely the Manually Labeled Multilingual Sentiment Analysis Dataset (ML-MSAD) and the Manually Labeled Emotion Analysis Multilingual Dataset (ML-EAD) are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABSA	Aspect-Based Sentiment Analysis
ABSAD	Aspect-Based Sentiment Analysis Dataset
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
COT	Chain-of-Thought
EA	Emotion Analysis
EAD	Emotion Analysis Dataset
EP	Emotional Prompting
ESAD	English Sentiment Analysis Dataset
GenAI	Generative Artificial Intelligence
GRU	Gated Recurrent Unit
LLM	Large Language Model
LSTM	Long Short-Term Memory
ML-EAD	Manually Labeled Emotion Analysis Multilingual Dataset

ML-MSAD	Manually Labeled Multilingual Sentiment Analysis Dataset
MSAD	Multilingual Sentiment Analysis Dataset
NA	Naming the Assistant
NLP	Natural Language Processing
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Pretraining Approach
RP	Role-Playing
SA	Sentiment Analysis
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TSD	Toxic Speech Detection

References

1. Fanni, S.C.; Febi, M.; Aghakhanyan, G.; Neri, E. Natural Language Processing. In *Imaging Informatics for Healthcare Professionals*; Springer International Publishing: Cham, Switzerland, 2023; p. 87. [\[CrossRef\]](#)
2. Qi, Y.; Shabrina, Z. Sentiment analysis using Twitter data: A comparative application of lexicon- and machine-learning-based approach. *Soc. Netw. Anal. Min.* **2023**, *13*, 31. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Kokab, S.T.; Asghar, S.; Naz, S. Transformer-based deep learning models for the sentiment analysis of social media data. *Array* **2022**, *14*, 100157. [\[CrossRef\]](#)
4. Areshey, A.; Mathkour, H. Exploring transformer models for sentiment classification: A comparison of BERT, RoBERTa, ALBERT, DistilBERT, and XLNet. *Expert Syst.* **2024**, *41*, e13701. [\[CrossRef\]](#)
5. Patil, R.; Gudivada, V.N. A Review of Current Trends, Techniques, and Challenges in Large Language Models (LLMs). *Appl. Sci.* **2024**, *14*, 2074. [\[CrossRef\]](#)
6. Vajjala, S.; Shimangaud, S. Text Classification in the LLM Era—Where do we stand? *arXiv* **2025**, arXiv:2502.11830. [\[CrossRef\]](#)
7. Hou, Z.; Zhao, R.; Li, Z.; Wang, Z.; Wu, Y.; Gou, J.; Zhu, Z.G. A Method for Efficient Structured Data Generation with Large Language Models. In *Proceedings of the 36th International Conference on Software Engineering and Knowledge Engineering*; Association for Computing Machinery: New York, NY, USA, 2024. [\[CrossRef\]](#)
8. Hutson, J.; Ratican, J. Advancing Sentiment Analysis Through Emotionally-Agnostic Text Mining in Large Language Models (LLMs). *J. Biosens. Bioelectron. Res.* **2024**, *1*, 118. [\[CrossRef\]](#) [\[PubMed\]](#)
9. He, L.; Yin, T.; Zheng, K. They May Not Work! An evaluation of eleven sentiment analysis tools on seven social media datasets. *J. Biomed. Inform.* **2022**, *132*, 104142. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Chowdhary, K.R. *Natural Language Processing*; Springer: New Delhi, India, 2020; p. 603. [\[CrossRef\]](#)
11. Kang, Y.; Cai, Z.; Tan, C.; Huang, Q.; Liu, H. Natural language processing (NLP) in management research: A literature review. *J. Manag. Anal.* **2020**, *7*, 139. [\[CrossRef\]](#)
12. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020. [\[CrossRef\]](#)
13. Islam, M.S.; Kabir, M.N.; Ghani, N.A.; Zamli, K.Z.; Zulkifli, N.S.A.; Rahman, M.M.; Moni, M.A. Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach. *Artif. Intell. Rev.* **2024**, *57*, 3. [\[CrossRef\]](#)
14. Wankhade, M.; Rao, A.C.S.; Kulkarni, C. A survey on sentiment analysis methods, applications, and challenges. *Artif. Intell. Rev.* **2022**, *55*, 5731. [\[CrossRef\]](#)
15. Mohd Nafis, N.S.; Awang, S. An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification. *IEEE Access* **2021**, *9*, 52177. [\[CrossRef\]](#)
16. Borkakoty, H.; Dev, C.; Ganguly, A. A Novel Approach to Calculate TF-IDF for Assamese Language. In *Lecture Notes in Electrical Engineering*; Springer: Berlin/Heidelberg, Germany, 2020; p. 387. [\[CrossRef\]](#)
17. Zhang, Q.; Xiao, J.; Tian, C.; Lin, J.C.; Zhang, S. A robust deformed convolutional neural network (CNN) for image denoising. *CAAI Trans. Intell. Technol.* **2022**, *8*, 331. [\[CrossRef\]](#)
18. Liu, Y.; Pu, H.; Sun, D. Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices. *Trends Food Sci. Technol.* **2021**, *113*, 193. [\[CrossRef\]](#)
19. Durga, P.; Godavarthi, D. Deep-Sentiment: An Effective Deep Sentiment Analysis Using a Decision-Based Recurrent Neural Network (D-RNN). *IEEE Access* **2023**, *11*, 108433. [\[CrossRef\]](#)

20. Das, S.; Tariq, A.; Santos, T.; Kantareddy, S.S.; Banerjee, I. Recurrent Neural Networks (RNNs): Architectures, Training Tricks, and Introduction to Influential Research. In *Neuromethods*; Humana Press: New York, NY, USA, 2023; p. 117. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Purwono, P.; Ma'arif, A.; Rahmiani, W.; Fathurrahman, H.I.K.; Frisky, A.Z.K.; Haq, Q.M.u. Understanding of Convolutional Neural Network (CNN): A Review. *Int. J. Robot. Control Syst.* **2023**, *2*, 739. [\[CrossRef\]](#)
22. Tembhurne, J.V.; Diwan, T. Sentiment analysis in textual, visual and multimodal inputs using recurrent neural networks. *Multimed. Tools Appl.* **2020**, *80*, 6871–6910. [\[CrossRef\]](#)
23. Li, Z.; Yang, C.; Huang, C. A Comparative Sentiment Analysis of Airline Customer Reviews Using Bidirectional Encoder Representations from Transformers (BERT) and Its Variants. *Mathematics* **2023**, *12*, 53. [\[CrossRef\]](#)
24. Gupta, P.; Gandhi, S.; Chakravarthi, B.R. Leveraging Transfer learning techniques- BERT, RoBERTa, ALBERT and DistilBERT for Fake Review Detection. In *Proceedings of the Forum for Information Retrieval Evaluation*; Association for Computing Machinery: New York, NY, USA, 2021; p. 75. [\[CrossRef\]](#)
25. Prytula, M. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of ukrainian language reviews. *Artif. Intell.* **2024**, *29*, 85. [\[CrossRef\]](#)
26. Liu, Z.; Yang, K.; Xie, Q.; Zhang, T.; Ananiadou, S. EmoLLMs: A Series of Emotional Large Language Models and Annotation Tools for Comprehensive Affective Analysis. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2024; p. 5487. [\[CrossRef\]](#)
27. Zhang, W.; Deng, Y.; Liu, B.; Pan, S.J.; Bing, L. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In *Findings of the Association for Computational Linguistics: NAACL*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3881–3906. [\[CrossRef\]](#)
28. Carneros-Prado, D.; Villa, L.; Johnson, E.; Dobrescu, C.C.; Barragán, A.; García-Martínez, B. Comparative Study of Large Language Models as Emotion and Sentiment Analysis Systems: A Case-Specific Analysis of GPT vs. IBM Watson. In *Lecture Notes in Networks and Systems*; Springer: Cham, Switzerland, 2023; p. 229. [\[CrossRef\]](#)
29. Zhang, B.; Yang, H.; Zhou, T.; Babar, M.A.; Liu, X.Y. Enhancing Financial Sentiment Analysis via Retrieval Augmented Large Language Models. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*; Association for Computing Machinery: New York, NY, USA, 2023; p. 349. [\[CrossRef\]](#)
30. Yadav, A. A Comprehensive Review on Large Language Models. In *Advances in Systems Analysis, Software Engineering, and High Performance Computing*; IGI Global: Hershey, PA, USA, 2024; p. 17. [\[CrossRef\]](#)
31. Liu, Y.; Zhu, X.; Shen, Z.; Liu, Y.; Li, M.; Chen, Y.J.; John, B.E.; Ma, Z.; Hu, T.; Li, Z.; et al. Do Large Language Models Possess Sensitivity to Sentiment? *arXiv* **2024**, arXiv:2409.02370. [\[CrossRef\]](#)
32. Kostina, A.; Dikaiakos, M.D.; Stefanidis, D.; Pallis, G. Large Language Models for Text Classification: Case Study and Comprehensive Review. *arXiv* **2025**, arXiv:2501.08457.
33. Go, A.; Bhayani, R.; Huang, L. *Twitter Sentiment Classification Using Distant Supervision*; CS224N Project Report; Stanford University: Stanford, CA, USA, 2009; pp. 1–12.
34. Morgan, Y. Twitter Tweets Sentiment Dataset. Kaggle. 2022. Available online: <https://www.kaggle.com/datasets/yasserh/twitter-tweets-sentiment-dataset> (accessed on 18 May 2026).
35. Barbieri, F.; Espinosa Anke, L.; Camacho-Collados, J. XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the 13th Language Resources and Evaluation Conference*; European Language Resources Association: Paris, France, 2022; p. 258.
36. Jp797498e. Twitter Entity Sentiment Analysis. Kaggle. Available online: <https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis> (accessed on 18 May 2026).
37. Elgiriye withana, N. Emotions. Kaggle, 2024. Available online: <https://www.kaggle.com/datasets/nelgiriye withana/emotions> (accessed on 18 May 2026). [\[CrossRef\]](#)
38. Namdari, R. Youtube Toxic Comments. Kaggle, 2024. Available online: <https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data> (accessed on 18 May 2026).
39. Gaeta, A.; Loia, V.; Lorusso, A.; Orciuoli, F.; Pascuzzo, A. Towards a LLM-based intelligent system for detecting propaganda within textual content. *Comput. Electr. Eng.* **2025**, *128*, 110765. [\[CrossRef\]](#)

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.