Multimodal Sentiment Analysis of Instagram Posts: A Comprehensive Practicum Report

Etana Disasa
Master of Science in Data Science
School of Computer and Information Sciences
Anderson College of Business and Computing, Regis University
edisasa@regis.edu

Table of Contents

Abstract

This practicum project developed an interactive multimodal sentiment analysis tool for Instagram posts, combining natural language processing (NLP) and computer vision techniques to analyze sentiment across images, captions, and comments. The system integrates multiple pre-trained models including BERT for text sentiment analysis, BLIP-2 for image description generation, YOLOv8 for face detection, and specialized models for emotion recognition and NSFW content filtering. The final deliverable is a Streamlit-based interactive application that accepts Instagram post URLs and provides comprehensive sentiment analysis with weighted scoring (50% media, 30% caption, 20% comments). The project addresses critical societal concerns about online toxicity and digital well-being while demonstrating practical applications in brand monitoring, mental health research, and social media analytics. Key contributions include the development of a reproducible multimodal analysis workflow, ethical AI deployment considerations, and a user-friendly interface for non-technical users.

1. Introduction and Background

1.1 Problem Context

Social media platforms, particularly Instagram, have fundamentally transformed digital communication by enabling users to share multimodal content combining images, videos, text captions, and comments. With over 2 billion active users, Instagram generates vast amounts of user-generated content daily, making it a critical platform for understanding public sentiment, emotional expression, and social interactions.

Traditional sentiment analysis approaches have predominantly focused on textual content, utilizing natural language processing techniques to classify emotions and opinions expressed in written form. However, this uni-modal approach fails to capture the rich, nuanced meaning conveyed through visual elements, which often constitute the primary communication medium on image-centric platforms like Instagram. Research has shown that visual content can significantly influence interpretation and emotional response, making integrated multimodal analysis essential for comprehensive sentiment understanding.

1.2 Societal Impact and Motivation

The proliferation of social media has coincided with increased concerns about digital well-being, online toxicity, and mental health implications of social media engagement. Studies have documented associations between social media use and negative outcomes including depression, anxiety, and harmful social comparison behaviors, particularly among younger demographics. The algorithmic amplification of engaging content, regardless of its emotional impact, has created environments where toxic interactions and negative sentiment can spread rapidly and influence user behavior.

This practicum addresses these concerns by developing tools that can identify and analyze sentiment patterns in social media content, potentially supporting interventions aimed at promoting healthier digital environments and mitigating the psychological harms associated with unregulated social media engagement.

1.3 Research Question and Objectives

The primary research question guiding this project was:

"How can a multimodal sentiment analysis system be developed to analyze Instagram post content (images, captions, comments) to provide sentiment insights that help identify harmful interactions?"

The project objectives included:

- Developing an integrated multimodal analysis pipeline combining computer vision and NLP techniques
- Creating an interactive, user-friendly application for real-time sentiment analysis
- Implementing ethical safeguards including NSFW content filtering and privacy protection
- Demonstrating practical applications for brand monitoring, research, and digital well-being initiatives
- Establishing a reproducible methodology for multimodal social media analysis.

2. Literature Review and Related Work

2.1 Multimodal Sentiment Analysis

The field of multimodal sentiment analysis has evolved significantly, with researchers recognizing the limitations of uni-modal approaches. Niu et al. (2016) introduced the MVSA dataset for visual sentiment analysis, demonstrating the feasibility of emotion classification from images using supervised learning methods. Tsai et al. (2019) advanced the field with multimodal transformer architectures that model cross-modal interactions, achieving state-of-the-art performance in sentiment classification tasks.

Comprehensive surveys by Poria et al. (2017) and Baltrusaitis et al. (2019) have systematically classified multimodal machine learning techniques, emphasizing their relevance for social media content analysis. These works highlight the importance of fusion strategies, feature alignment, and cross-modal learning in developing effective multimodal systems.

2.2 Social Media and Mental Health

Research examining the psychological impacts of social media engagement has revealed concerning patterns. Lup et al. (2015) identified associations between Instagram use, negative social comparison, and depressive symptoms, particularly among young adults. Verduyn et al. (2017) conducted comprehensive reviews demonstrating that passive social media consumption often undermines subjective well-being and contributes to negative mood states.

The emergence of online toxicity has been examined by Cheng et al. (2017), who showed how ordinary users can become trolls under certain social conditions, contributing to harmful online environments. These findings underscore the importance of developing tools that can identify and potentially mitigate negative interactions in social media contexts.

2.3 Ethical AI and Algorithmic Accountability

Critical perspectives on social media algorithms, exemplified by documentaries like "The Social Dilemma" (Orlowski, 2020), have exposed how engagement optimization algorithms exploit human psychological vulnerabilities to maximize attention and platform usage. Noble's (2018) "Algorithms of Oppression" demonstrates how algorithmic systems can reinforce societal biases and discrimination.

Ethical AI frameworks such as AI4People (Floridi et al., 2018) advocate for designing AI systems that prioritize societal good, transparency, and harm reduction. Chancellor and De Choudhury (2020) provide critical reviews of predictive mental health inference methods, emphasizing risks of misuse, privacy violations, and diagnostic overreach in automated systems.

3. Methodology and Technical Architecture

3.1 Methodology Evolution and Scope Adjustment

The original proposal outlined a plan to fine-tune custom models on publicly available datasets including Sentiment140, GoEmotions for text, and the FI Dataset for images. However, after careful assessment of the practicum timeline and scope constraints, the methodology was strategically adjusted to leverage existing pre-trained models. This pivot allowed for:

- More comprehensive focus on multimodal integration and system architecture
- Enhanced attention to ethical considerations and user experience design
- Broader exploration of different model combinations and fusion strategies
- Development of a fully functional application within the practicum timeframe

This methodological evolution from custom model training to pre-trained model integration represents a pragmatic approach to project management while maintaining the core research objectives of multimodal sentiment analysis.

3.2 System Architecture Overview

The multimodal sentiment analysis system was designed as a four-stage pipeline:

1. **Data Ingestion**: Instagram post data retrieval and validation
2. **Media Analysis**: Computer vision processing for images and videos
3. **Text Analysis**: Natural language processing for captions and comments
4. **Sentiment Consolidation**: Weighted fusion of multimodal sentiment scores

The system architecture emphasizes modularity, with clear separation between the main Streamlit application (sentiment-processor.py) and utility functions (socialmedia_utils.py). This design enables maintainable code and efficient resource management through strategic model caching.

3.3 Data Ingestion Module

The data ingestion process utilizes Instaloader, a Python library for accessing Instagram post data. The system accepts Instagram post URLs, validates the links, and extracts publicly available content including:

- Images and video frames
- Post captions
- Top 5 comments (ranked by engagement)
- Metadata (likes, post type, number of media items)

The implementation includes comprehensive error handling mechanisms that ensure graceful failure when invalid URLs are provided or when posts are inaccessible due to privacy settings. The system uses session management with pre-configured username credentials to access Instagram's API through Instaloader.

Error handling mechanisms ensure graceful failure when invalid URLs are provided or when posts are inaccessible due to privacy settings.

3.4 Media Analysis Pipeline

The media analysis component implements several computer vision models:

NSFW Content Filtering: A pre-trained NSFW detection model screens all media content, filtering inappropriate material to ensure ethical deployment and prevent misuse of the system.

Image Description Generation: BLIP-2 (Bootstrapped Language-Image Pretraining) generates descriptive captions for images and video frames, creating textual representations of visual content that can be processed by NLP sentiment analysis models.

Face Detection and Emotion Analysis: YOLOv8 detects faces in images with a confidence threshold of 70% or higher. Detected faces are processed by a specialized face emotion recognition model that classifies expressions into standard emotion categories.

Video Processing Enhancement: The media analysis component evolved during development to include sophisticated video processing. Initially, video content was not considered for analysis, then basic thumbnail extraction was implemented, and finally the system was enhanced to extract specific frames from within videos for comprehensive analysis. OpenCV extracts representative frames from videos at configurable positions (default: 30% through the video timeline), treating video content as sequences of analyzable static images. This iterative improvement significantly expanded the system's capability to handle Instagram's diverse media types.

3.5 Text Analysis Pipeline

The text processing component employs several NLP techniques:

BERT-based Sentiment Classification: Pre-trained BERT models (specifically bhadresh-savani/bert-base-go-emotion) analyze sentiment in captions, comments, and generated image descriptions. This model provides emotion classification across multiple categories including joy, sadness, anger, fear, surprise, and others.

Emoji Processing: The emoji library converts emoji characters to their textual descriptions before sentiment analysis, ensuring that emotional expressions conveyed through emojis are captured in the analysis.

Comment Ranking: The system analyzes the top 5 comments based on engagement metrics, focusing on content most likely to influence public perception and sentiment.

3.6 Sentiment Consolidation and Integration

The final sentiment score is calculated using a weighted combination approach:

- **Media Sentiment (50% weight)**: Analysis of image descriptions and face emotions detected in visual content
- **Caption Sentiment (30% weight)**: Direct sentiment analysis of post caption
- **Comment Sentiment (20% weight)**: Analysis of available comments when accessible
- **Overall summary**: Statistical aggregation of all detected emotions with visualization

This weighting scheme reflects the visual-centric nature of Instagram while acknowledging the importance of textual context and community engagement through comments.

4. Tools and Libraries

4.1 Core Libraries and Frameworks

Data Retrieval and Processing:

- Instaloader: Instagram post data collection
- OpenCV: Video frame extraction and image processing
- Pandas, NumPy: Data manipulation and analysis

Computer Vision:

- YOLOv8: State-of-the-art object detection for face identification
- BLIP-2: Advanced image captioning using vision-language models
- Custom NSFW detection model: Content filtering for ethical deployment
- Specialized face emotion recognition model

Natural Language Processing:

- Transformers library: BERT model implementation
- Emoji library: Emoji-to-text conversion for sentiment analysis
- Pre-trained BERT models from Hugging Face model hub

Application Development and Deployment:

- Streamlit: Interactive web application framework
- Matplotlib/Seaborn: Data visualization and results presentation

4.2 Model Selection Rationale

The selection of pre-trained models was based on several criteria:

- Performance: Models demonstrated strong performance on benchmark datasets
- Efficiency: Models could run in reasonable time for real-time applications
- Availability: Models were publicly available and well-documented
- Ethical considerations: Models had been developed with responsible AI principles.

5. Implementation and Results

5.1 Application Interface and User Experience

The Streamlit-based application provides an intuitive interface designed for both technical and non-technical users. Key features include:

URL Input and Validation: Users enter Instagram post URLs, with immediate validation and error messaging for invalid inputs.

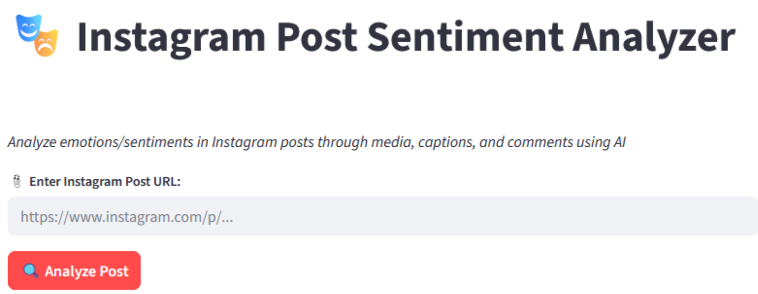

Image 1. Streamlit Application

Content Overview: The application displays post metadata including media type (image/video), number of comments, likes, and caption excerpts. A post from Regis University's official Instagram account was utilized (https://www.instagram.com/p/DNn_uSsx5xJ/).

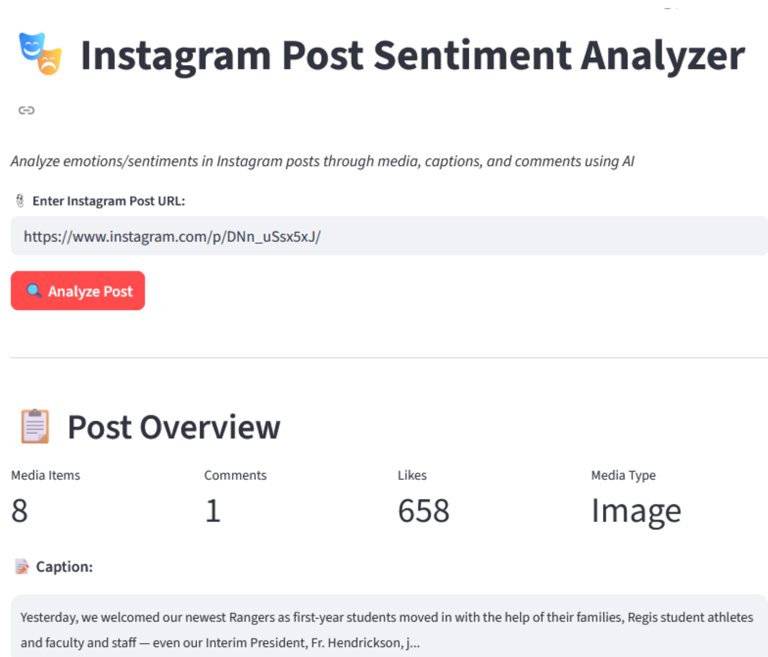

Image 2. A Post Overview section

Media Analysis Results: For each piece of media, the system displays:

- Thumbnail images with unique identifiers
- AI-generated descriptions
- Sentiment scores from image analysis
- Face detection results with emotion classifications and confidence scores
- Detected faces will be displayed in a dedicated expandable section

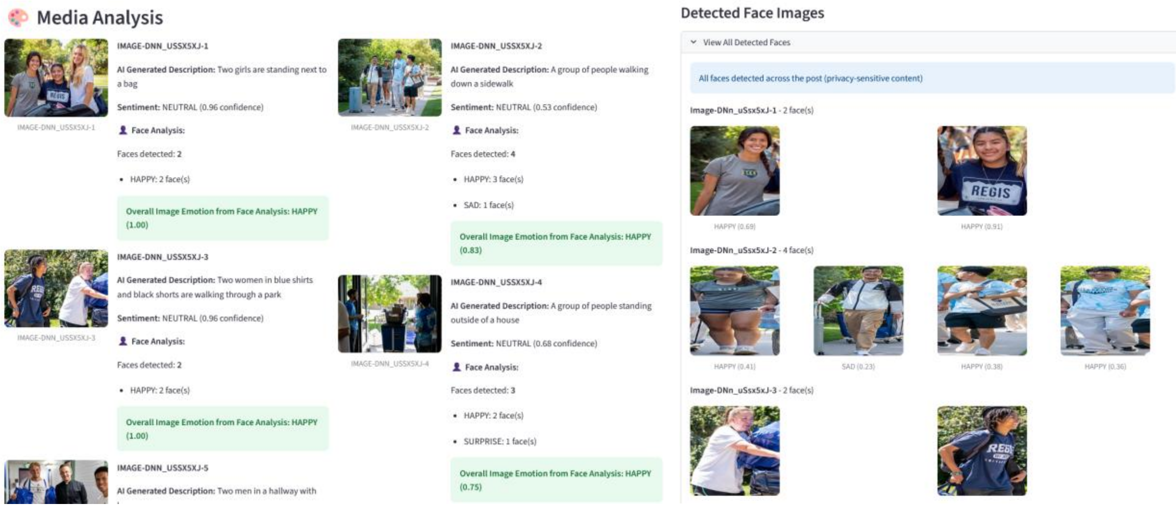

Image 3. Media Analysis section displays media including faces identified

Text Analysis Results: Caption and comment sentiment analysis results are presented with clear sentiment classifications and confidence scores.

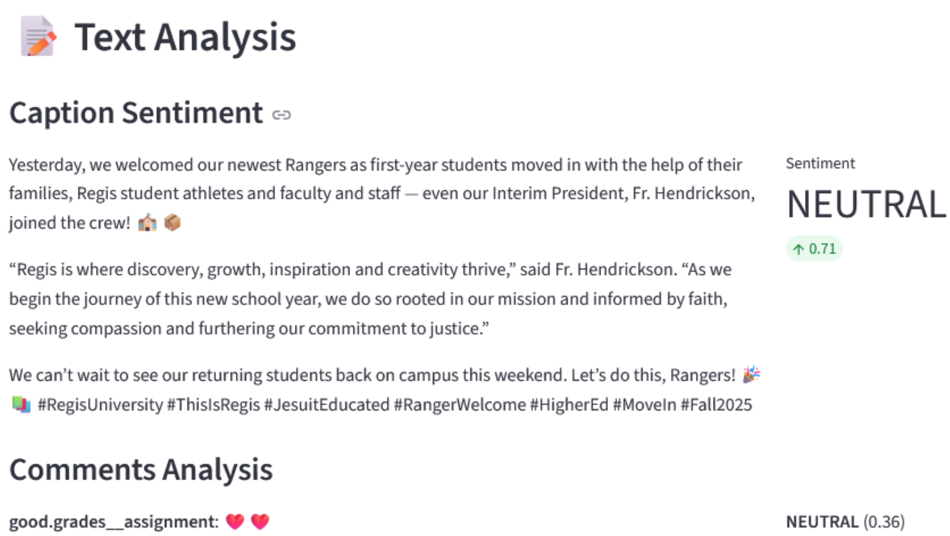

Image 4. Text analysis shows both caption sentiment and comment sentiment

Final Summary Visualization: A comprehensive summary presents the weighted overall sentiment using intuitive color schemes and graphical representations.

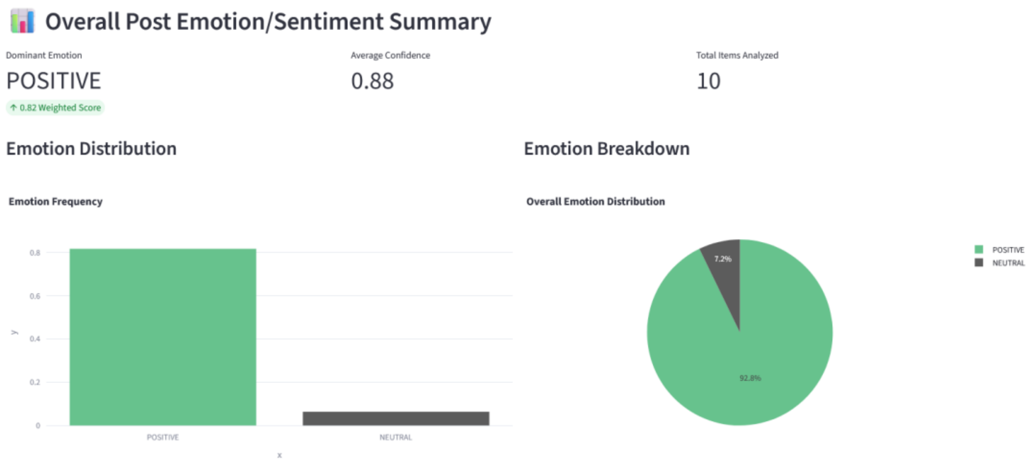

Image 5. Over all post sentiment is calculated and displayed

5.2 Application Performance and Content Filtering

The system was tested on various Instagram posts, including on several Regis University posts. Analysis of an official university posts demonstrated the system's ability to:

- Process institutional content with appropriate imagery
- Generate accurate descriptions of educational/promotional content
- Analyze professional captions and positive community engagement
- Provide balanced sentiment scores reflecting the content's positive nature

Multiple posts from news agencies, prominent individuals, and large accounts were tested; in some cases, the NSFW mechanism successfully flagged inappropriate content, confirming system reliability. For privacy reasons, links to these posts are not included in this report

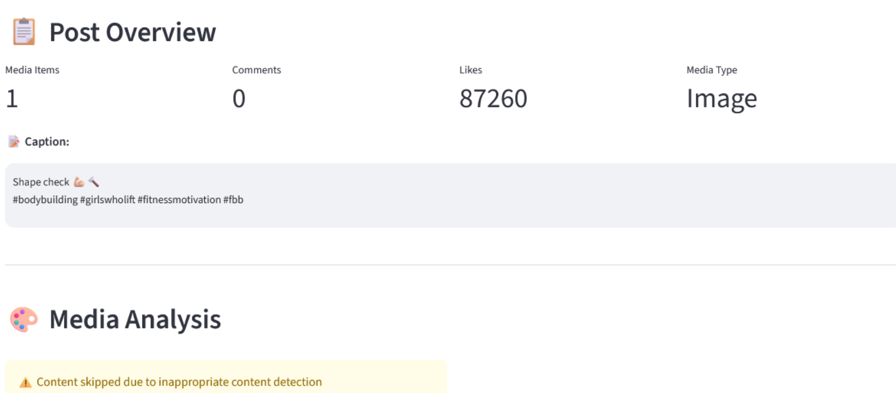

Image 6. System response to NSFW content: flagged media and interrupted processing.

5.3 Evaluation Approach and Scope Limitations

Qualitative Assessment Over Quantitative Metrics: Given the strategic pivot to pre-trained models rather than custom model development, the evaluation focused on qualitative assessment and functional validation rather than traditional machine learning metrics such as F1-scores, precision, and recall.

This approach was appropriate for several reasons.

Integration Success Metrics: The primary achievement lay in successfully integrating multiple state-of-the-art pre-trained models into a cohesive multimodal analysis pipeline, which was evaluated through:

- Successful processing of diverse Instagram content types
- Stable performance across different post formats (single images, carousels, videos)
- Appropriate handling of edge cases and error conditions
- User interface responsiveness and clarity

Functional Validation: Rather than training custom models that would require traditional ML evaluation metrics, the project demonstrated:

- Effective multimodal fusion of computer vision and NLP outputs
- Robust content filtering and safety measures
- Intuitive user experience design for non-technical users
- Ethical AI implementation with privacy protections

Scope-Appropriate Evaluation: The evaluation methodology was designed to align with the project's progression from model exploration to full system integration, emphasizing practical utility and real-world applicability. This approach reflected a balanced scope, prioritizing functional deliverables and broader relevance over purely academic benchmarks that would have required extensive model training and validation datasets.

6. Ethical Considerations and Responsible AI Implementation

6.1 Privacy and Consent

The system was designed with strong privacy protections:

- Analysis limited to publicly available Instagram posts only
- No storage of personal data or post content beyond session duration
- No access to private messages or restricted content
- Clear communication about data usage and limitations

6.2 Content Sensitivity and Safety

Multiple safeguards address content sensitivity:

- NSFW filtering prevents analysis of inappropriate material
- Face detection confidence thresholds (≥70%) reduce false positives
- Clear labeling of AI-generated content and automated analysis results
- Transparency about model limitations and potential biases

6.3 Responsible Use and Limitations

The system includes clear documentation about appropriate use:

- Analysis results are informational, not diagnostic
- Not intended for individual profiling or discriminatory actions
- Designed for research, monitoring, and educational purposes
- Acknowledgment of potential cultural, gender, and racial biases in pre-trained models

6.4 Transparency and Interpretability

The system provides transparent analysis processes:

- Clear documentation of all models and preprocessing steps
- Confidence scores and uncertainty indicators for all predictions
- Visual presentation of intermediate results (face detection, image descriptions)
- Open acknowledgment of system limitations and potential errors

7. Challenges and Limitations

7.1 Technical Challenges

- **Python Compatibility Issues:** DeepFace was incompatible with Python 3.13.7, prompting a pivot to an alternative emotion detection setup using dima806/facial_emotions_image_detection and YOLOv8 for face localization.
- **Video Frame Extraction:** Instagram's default thumbnails poorly represented video content. Reliable frame selection required custom OpenCV logic to extract frames at a more representative point (e.g., 30% into the video).
- **Session Management with Instagram:** Instaloader sessions initially failed due to browser-based session expiry. Using persistent sessions via terminal login (instaloader --login) resolved the issue and reduced API disruptions.
- **YOLOv8 Integration:** Limited documentation and unclear repo structure made integrating YOLOv8 and identifying the correct model weights (e.g., yolov8n.pt) more time-consuming than expected.
- **NSFW Detection Constraints:** Instagram's content moderation limited exposure to explicit material, restricting real-world testing of the NSFW classifier.

- **Performance Bottlenecks:** Frequent API calls led to session blocking. Caching strategies and optimized model loading were required to maintain system responsiveness and avoid repeated loading overhead.

7.2 Methodological Limitations

- **Library Compatibility:** Coordinating multiple libraries (e.g., PyTorch, OpenCV, Transformers) under Python 3.13.7 required careful dependency management to avoid conflicts.
- **Resource Constraints:** Loading multiple large models (BLIP, BERT, YOLOv8, emotion detectors) taxed memory and startup times. Streamlit caching helped, but performance remained a challenge.
- **Evaluation Trade-offs:** Relying on pre-trained models meant traditional metrics like precision and recall were not applicable. Evaluation focused instead on functional validation and user experience.

7.3 Data & Content Limitations

- **Limited Access Scope:** Only public Instagram posts were analyzed, excluding private content where toxic or sensitive behavior may be more prevalent.
- **Temporal Blind Spots:** The system provides a snapshot analysis, without tracking how sentiment or context evolves over time.
- **Bias in Pre-Trained Models:** Cultural and linguistic biases from Western-trained models may reduce accuracy for non-English or culturally specific content.
- **Comment Sampling Limits:** Only the top 5 most-engaged comments per post were analyzed, which may not reflect broader sentiment trends or community nuance.

7.4 Ethical Considerations

- **Over-Reliance Risk:** Automated sentiment assessments may be misinterpreted as objective truth, despite their limitations in emotional nuance.
- **Privacy Implications:** Though restricted to public content, face detection and emotion analysis raise ethical concerns around consent and inferred emotional states.

8. Contributions and Significance

8.1 Technical Contributions

Multimodal Integration: Demonstrated successful integration of state-of-the-art computer vision and NLP models in a unified analysis pipeline, contributing to the growing field of multimodal AI applications.

Reproducible Methodology: Established a clear, reproducible workflow for multimodal social media analysis that can be adapted for other platforms and use cases.

Ethical AI Implementation: Incorporated comprehensive ethical considerations and safety measures, providing a template for responsible AI deployment in social media analysis.

User-Centered Design: Created an accessible interface enabling non-technical users to benefit from advanced AI capabilities, bridging the gap between research and practical application.

8.2 Societal and Research Contributions

Digital Well-being Tools: Contributed to the development of tools that can support digital well-being initiatives and interventions aimed at promoting healthier online environments.

Toxicity Detection: Advanced capabilities for identifying potentially harmful content and interactions, supporting efforts to combat online toxicity and cyberbullying.

Brand and Research Applications: Provided practical tools for brand monitoring, market research, and academic studies of social media behavior and sentiment trends.

Educational Impact: Demonstrated integration of multiple AI disciplines (computer vision, NLP, human-computer interaction) in a single project, providing educational value for data science curricula.

8.3 Broader Impact on Multimodal AI Research

Cross-Modal Understanding: Contributed to research demonstrating how different modalities (visual, textual) can be effectively combined for richer understanding of human communication and emotional expression.

Real-World Application: Provided a concrete example of how academic multimodal AI research can be translated into practical tools with societal relevance.

Ethical Framework: Established patterns for incorporating ethical considerations into multimodal AI systems, contributing to responsible AI development practices.

9. Future Work and Extensions

9.1 Technical Enhancements

- **Multilingual Support:** Adding multi-language capabilities would broaden global applicability and research relevance.
- **Temporal Analysis:** Integrating time-series analysis to track sentiment trends over time.

- **Cross-Platform Integration:** Extending the system to other platforms (e.g., Twitter, TikTok, Facebook) to enhance utility and reach.
- **Advanced Fusion Techniques:** Exploring more sophisticated multimodal fusion methods, such as attention mechanisms and neural architectures.

9.2 Evaluation and Validation

- **Custom Model Development:** Training domain-specific models could improve Instagram-specific accuracy and enable use of standard ML metrics.
- **Comparative Analysis:** Benchmarking against other sentiment analysis systems using standardized evaluation protocols.
- **User Studies:** Conducting formal UX studies to refine interface design and result interpretation.
- **Bias Assessment:** Analyzing model biases and developing strategies to ensure fair performance across diverse populations.
- **Longitudinal Studies:** Evaluating the system's impact and reliability over extended real-world usage.

9.3 Application Extensions

- **Mental Health Research:** Collaborating with researchers to explore digital mental health monitoring and therapeutic applications.
- **Crisis Detection:** Building features to detect potential self-harm or mental health crises in social media content.
- **Brand Intelligence:** Applying the system to track brand sentiment, customer feedback, and marketing effectiveness.
- **Educational Applications:** Creating educational tools to teach multimodal AI and responsible AI deployment.

10. Conclusion

This practicum successfully developed a multimodal sentiment analysis system for Instagram, integrating computer vision and natural language processing to analyze posts while addressing digital well-being and online toxicity. The project coordinated multiple state-of-the-art AI models, established a reproducible methodology, and delivered a user-friendly interface, making advanced sentiment analysis accessible to non-technical users.

The system's weighted fusion approach (50% media, 30% captions, 20% comments) balances visual and textual context, reflecting Instagram's content dynamics. Ethical considerations, including NSFW filtering, privacy protections, and transparency measures, provide a template for responsible AI deployment.

Practical applications span brand monitoring, academic research, and digital well-being initiatives. While limitations exist in evaluation scope, language support, and potential model biases, the project lays a foundation for future enhancements such as multilingual support, temporal analysis, and cross-platform integration.

Overall, this work demonstrates how academic research in multimodal AI can be translated into practical, socially valuable tools. The experience gained—technical integration, ethical implementation, and UX design—offers insights for future practitioners at the intersection of AI and social media analysis. As social media continues to shape communication, tools like this system will be increasingly important for understanding and improving digital environments.

References

Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *npj Digital Medicine*, 3(1), 1–11.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017). Antisocial behavior in online discussion communities. *ICWSM*, 11, 61–70.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.

Kross, E., Verduyn, P., Demiralp, E., Park, J., Lee, D. S., Lin, N., ... & Ybarra, O. (2021). Social media and well-being: Pitfalls, progress, and next steps. *Trends in Cognitive Sciences*, 25(1), 55–66.

Lup, K., Trub, L., & Rosenthal, L. (2015). Instagram instasad?: Exploring associations among Instagram use, depressive symptoms, negative social comparison, and strangers followed. *Cyberpsychology, Behavior, and Social Networking*, 18(5), 247–252.

Niu, X., Li, H., Xu, X., & Xu, C. (2016). Visual sentiment analysis on Flickr images. *IEEE Transactions on Multimedia*, 18(11), 2224–2233.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Orlowski, J. (Director). (2020). *The Social Dilemma* [Film]. Netflix.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37, 98–125.

Tsai, Y. H. H., Bai, S., Yamada, M., Morency, L. P., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 6558–6569.

Verduyn, P., Ybarra, O., Resibois, M., Jonides, J., & Kross, E. (2017). Do social network sites enhance or undermine subjective well-being? A critical review. *Social Issues and Policy Review*, 11(1), 274–302.