TYPE Original Research
PUBLISHED 24 October 2024
DOI 10.3389/fcomp.2024.1444549

Check for updates

OPEN ACCESS

EDITED BY
Rocco Pietrini,
Marche Polytechnic University, Italy

REVIEWED BY
Lorenzo Stacchio,
University of Macerata, Italy
Pasquale Cascarano,
University of Bologna, Italy

*CORRESPONDENCE
Jon Chun
✉ chunj@kenyon.edu

RECEIVED 10 July 2024
ACCEPTED 17 September 2024
PUBLISHED 24 October 2024

CITATION
Chun J (2024) MultiSentimentArcs: a novel method to measure coherence in multimodal sentiment analysis for long-form narratives in film.
Front. Comput. Sci. 6:1444549.
doi: 10.3389/fcomp.2024.1444549

MultiSentimentArcs: a novel method to measure coherence in multimodal sentiment analysis for long-form narratives in film

Jon Chun*

KDHLab, Integrated Program for Humane Studies, Kenyon College, Gambier, OH, United States

Affective artificial intelligence and multimodal sentiment analysis play critical roles in designing safe and effective human-computer interactions and are in diverse applications ranging from social chatbots to eldercare robots. However emotionally intelligent artificial intelligence can also manipulate, persuade, and otherwise compromise human autonomy. We face a constant stream of ever more capable models that can better understand nuanced, complex, and interrelated sentiments across different modalities including text, vision, and speech. This paper introduces MultiSentimentArcs, combination of an open and extensible multimodal sentiment analysis framework, a challenging movie dataset, and a novel benchmark. This enables the quantitative and qualitative identification, comparison, and prioritization of conflicting sentiments commonly arising from different models and modalities. Diachronic multimodal sentiment analysis is especially challenging in film narratives where actors, directors, cinematographers and editors use dialog, characters, and other elements in contradiction with each other to accentuate dramatic tension. MultiSentimentArcs uses local open-source software models to democratize artificial intelligence. We demonstrate how a simple 2-step pipeline of specialized open-source software with a large multimodal model followed by a large language model can approximate video sentiment analysis of a commercial state-of-the-art Claude 3 Opus. To the best of our knowledge, MultiSentimentArcs is the first fully open-source diachronic multimodal sentiment analysis framework, dataset, and benchmark to enable automatic or human-in-the-loop exploration, analysis, and critique of multimodal sentiment analysis on long-form narratives. We demonstrate two novel coherence metrics and a methodology to identify, quantify, and explain real-world sentiment models and modalities. MultiSentimentArcs integrates artificial intelligence with traditional narrative studies and related fields like film, linguistic and cultural studies. It also contributes to eXplainable artificial intelligence and artificial intelligence safety by enhancing artificial intelligence transparency in surfacing emotional persuasion, manipulation, and deception techniques. Finally, it can filter noisy emotional input and prioritize information rich channels to build more performant real-world human computer interface applications in fields like e-learning and medicine. This research contributes to the field of Digital Humanities by giving non-artificial intelligence experts access to directly engage in analysis and critique of research around affective artificial intelligence and human-AI alignment. Code and non-copyrighted data will be available at https://github.com/jon-chun/multisentimentarcs.

KEYWORDS

AI, multimodal sentiment analysis, diachronic sentiment analysis, open-source AI, LLM, LMM, narrative, storytelling

1 Introduction

Rapid progress in artificial intelligence (AI) has led to remarkable breakthroughs in tasks involving reasoning, language, and emotion. AI has achieved human-level performance across a wide range of deterministic, logical reasoning and skill-based tasks (Achiam et al., 2023). State-of-the-art (SOTA) AI models have even demonstrated superhuman performance and unexpected emergent capabilities (Bubeck et al., 2023). For these reasons, static AI benchmarks are rapidly becoming obsolete (Lynch, 2023) and large-scale evolving benchmarks like BIG-Bench (Srivastava, 2022) and HELM (Liang et al., 2023) attempt to remain relevant by aggregating hundreds of existing and new individual metrics.

Affective Computing develops systems capable of recognizing, interpreting, processing, and simulating human emotions (Picard, 1997; Blanchard et al., 2009). Despite recent breakthroughs leveraging progress in AI, sometimes termed affective AI, the field faces many challenges (Afzal et al., 2023; Zhang et al., 2024). Emotions are often personal, culturally and temporally relative, and lacking reliable, coherent, and well-defined ground truth references (Altarriba and Kazanas, 2017; Quesque et al., 2022; Saxena et al., 2020; Brooks et al., 2024). To further complicate things, emotions are mediated through underspecified modalities like vision and language. Meaning is often implied or even in conflict with other modalities (Dylman et al., 2020). The emerging field of machine psychology often uses psychometric tests designed for humans in affective AI studies (Wang, 2022), but there are few techniques developed specifically for AI. These gaps in multimodal SA represent a growing AI Safety threat. This reality is reflected in the EU AI Act, which classifies human emotion recognition as high risk given the potential to manipulate and undermine human autonomy (EU Commission, 2024). There is growing interest in particular risks related to persuasion and deception (Salvi et al., 2024), yet the intersection between AI safety, human-AI alignment and affective AI are relatively unexplored (Latif, 2022).

Historically, Good Old-Fashioned AI (GOFAI) has been constrained by the limitations of traditional approaches that include rule-based expert systems, expert feature engineering, and statistical machine learning (ML). These paradigms suited computationally tractable, narrow, well-defined, and deterministic applications like medical expert systems (Abu-Nasser, 2017). GOFAI generally lacked the ability to express, much less manipulate or optimize, everyday non-tabular tasks involving vision, language, and emotion. The current AI Spring based upon deep neural networks began with the 2012 ImageNet competition (Krizhevsky et al., 2012) and accelerated with the 2017 Transformers Google research paper (Vaswani, 2017). AI was brought to the wider public with the release of ChatGPT in late 2022. ChatGPT excelled at tasks on fuzzy data like language, vision, and complex reasoning by learning implicit statistical latent patterns in an immense training dataset. Under this new paradigm, AI engineers with big data, large scale compute, and refinements in architecture, algorithms, and training have replaced domain experts who traditionally carefully engineered features (Sutton, 2019).

Transformer-based large language models (LLMs) and their multimodal variants may have key limitations in functionality like world modeling (Murphy and Criddle, 2023), complex reasoning (Patel, 2024), and modeling causality (Ashwani et al., 2024). Nonetheless, advances in scale and training have led to superhuman performances on both narrow tasks (CITE) and emergent abstract abilities like theory of mind (ToM)—the ability to reason about the mental and emotional state of other core to human social intelligence (Street, 2024; Strachan et al., 2024). More significantly, humans have shown a preference for the emotionally intelligent chatbot Xiaoice over humans (Zhou et al., 2020) and generate 20 k hits/s at character. ai equaling 20% of Google search traffic (Chung, 2024).

Humans often view themselves as rational beings. The rise of behavioral economics was in response to the traditional concept of homo economicus, rational actors ideally self-optimizing all decisions with perfect information to create efficient markets (Thaler, 2015). Cognitive science reveals that human reason may be somewhat illusory and even often a soothing narrative we create post-hoc to preserve and protect our self-identity (Lavazza and Inglese, 2023). Johnathan Haidt gives the mental model of a small man (consciousness) atop a massive elephant (unconscious) struggling to control and direct its movement (Burke, 2012).

In contrast, a wide range of fields with real-world consequences are premised on connecting, persuading, and manipulating this elephant instead of the rider. Applications include sales, advertising, marketing, political campaigns, consoling/psychotherapy, propaganda/brainwashing, teambuilding/elite training, magic, etc. Our own lack of self-awareness of our numerous cognitive biases, social manipulations, and evolutionary exploits exposes an immense attack surface for superhuman affective AI. AI is ever more performant, aided by growing streams of both general and personalized high-quality data exhaust: email, internet, smart TV, mobile phone, credit card, EV car, smart home, and wearable IoTs like health watches or AR/VR glasses. The combination of rapid progress in AI, human vulnerabilities and the commercial incentives to exploit creates an urgent need to better understand, critique and design emotionally intelligent AI systems.

In this paper, we introduce MultiSentimentArcs (MultiSA), a multimodal affective AI framework that builds upon text-based SentimentArcs (Chun, 2023) and enables human-in-the-loop (HIL) exploration (Mosqueira-Rey et al., 2022), interpretation, and explainability of sentiment arcs in long-form narrative. Our approach leverages diachronic sentiment analysis techniques (Elkins, 2022; Fernández-Cruz and Moreno-Ortiz, 2023, Chun and Elkins, 2023) that generalize to various narrative forms, including novels, social media discussions, conversations, and medical narratives. By incorporating SOTA commercial APIs as well as open-source software (OSS) models that can run on mid-range consumer hardware, we aim to democratize access to these powerful analytical tools. HIL analysis enables researchers and practitioners to better understand, interpret, and align the behavior of affective AI systems, thereby mitigating the risks they pose to human agency and decision-making.

This paper makes the following contributions to help redress this deficit:

- To the best of our knowledge, MultiSA is the first fully **OSS multimodal affective AI framework** with extensible parallel pipelines to enable HIL exploration, interpretation, and explainability in long-form narratives. This includes movies, TV episodes, social media feeds, news, and debates based upon a proven method of close and distant reading adapted from literary analysis (Elkins, 2022).
- Depending upon the application, multimodal sentiment analysis (MSA) often diverges in sentiment agreement across modalities

from high (clinical with physiological signals) to low (film with dramatic and artistic choices to create tension). We present **two novel diachronic MSA distance metrics** to quantify the coherence of sentiment time series in both absolute and relative similarities terms using using Euclidean and dynamic time warping (DTW) techniques. This enables MSA systems to identify, quantify, and prioritize often conflicting sentiments for greater transparency and better performance.

- We present a novel inexpensive two step pipeline using an OSS large multimodal model (LMM) followed by a large language model (LLM) to perform diachronic sentiment analysis on video and compare it against results from a SOTA commercial LMM in Claude 3 to benchmark the **viability of OSS LLM/LMM alternatives** to MSA.
- Our benchmark dataset consists of 66 classic films from the golden-era of Hollywood across 8 genres including adventure, comedy, drama, film noir, musicals, psychological-thrillers, and westerns. This is an especially **challenging diachronic MSA dataset** for MSA applications given the high disagreement between models and modalities due to artistic, cinematographic and directorial storytelling techniques.
- We aid in the **democratization of AI** by incorporating leading OSS video processing and multimodal LLMs that can be run locally on mid-range consumer gaming laptops. This enables academics and narrative experts beyond traditional AI fields to lend their expertise to analyze, critique and generate annotated datasets to guide AI safety training and human-AI alignment.

2 Background

Affective AI is intrinsically multidisciplinary and expensive, but this section will focus on a few fields directly relevant to understanding MultiSA.

2.1 Language and meaning

Language has been central to the history of "Artificial Intelligence" since the term was coined at the Dartmouth conference in 1956. At the height of the Cold War, Russian-English machine translation was one of the first applications of AI. It was initially viewed as an easy task based upon how minimally computationally demanding it was for humans. By the 1980s, Moravec's Paradox formalized the idea that cognitive/perception tasks that are easy for humans like vision and language were difficult for computers. The reverse is also true: large scale complex calculations that are trivial for computers can be difficult or impossible for humans (Agrawal, 2010).

Linguistics traditionally describes language as operating on three levels: syntax, semantics, and pragmatics. Informally, syntactic structures are the physical manifestations of language: printed symbols, words, and grammatical rules to organize them. Semantics is the meaning behind the words, which is sometimes at odds with syntax and is reflected in challenging narrow natural language processing (NLP) tasks like humor, irony and comedy. Finally, pragmatics is the larger context within which language operates and is given meaning and interpreted beyond syntax or semantics.

Deep Neural Networks (DNNs) and LLMs finally succeeded in AI NLP after decades of failure from earlier tools and formalisms like LISP and Chomsky's hierarchies (Evans and Levinson, 2009). Contrary to expectations, carefully hand-engineered features and extensive rulesets were unable to capture natural language. Simply training large scale artificial neural networks on billions and trillions of tokens from the Internet and other sources of natural language worked better (Minaee, 2024). The scale of LLMs makes them largely impenetrable black boxes, however. This raises concerns about transparency, AI safety (NIST, 2023) and human-AI alignment (Anthropic, 2024b). HIL exploration, critique, and interpretation is one technique to build trust and ensure safety of AI systems.

2.2 Sentiment analysis

"Sentiment Analysis," also known as Opinion Mining, originated as a narrow NLP task that involves classifying emotions and quantifying negative/positive valence or polarity (Birjali et al., 2021). The two most popular emotional taxonomies classify emotions into distinct categories like Ekman's six fundamental emotions: joy, surprise, sadness, fear, disgust, and anger (Ekman et al., 1985) or Plutchik's wheel of 8 primary bipolar emotions (joy versus sadness; anger versus fear; trust versus disgust; and surprise versus anticipation; Plutchik, 1980). In this preliminary study, we restrict our sentiment analysis to only the positive or negative valence or polarity rather than labeling each emotion.

Different approaches to textual sentiment analysis include lexicons of words assigned fixed values (Hu and Liu, 2004), rules that modify these lexical values (Hutto and Gilber, 2014), statistical ML classifiers like Naïve Bayes (Loria, 2020), fine-tuned BERT classifiers (Romero, 2024), and more general purpose LLMs like OpenAI GPT4o and Anthropic Claude 3.0 Opus (Momennejad, 2023). Simpler models offer speed, cost, and portability at the cost of performance (Mabrouk, 2020). Fine-tuned small BERT model variants dominate the Sentiment Analysis Leaderboards (Papers with Code, 2024) while SOTA commercial LLMs do well by conventional metrics (Zhang, 2023) and can provide additional functionality like providing explanations (Krugmann and Hartmann, 2024).

Diachronic sentiment analysis tracks the changes of sentiment over time. For a written narrative like a novel or screenplay, the text is segmented into smaller sequential chunks like sentences, and the sentiment of each text chunk is used to form a numerical time series that is smoothed to visualize an emotional arc. Four major foundational works include "The emotional arcs of stories are dominated by six basic shapes" (Reagan, 2016), the Syuzhet.R library (Jockers, 2019), the Sentimentr.R library (Rinker, 2021), and SentimentArcs, a comprehensive ensemble in Python with a methodology based in literary analysis published by Cambridge UP (Elkins, 2022; Chun, 2021).

Reagan analyzed 1,327 stories from Gutenberg.org and found the smoothed sentiment arcs could be clustered into one of six archetype shapes like "rags to riches" or "man on a hill." This was a computational realization of Kurt Vonnegut's rejected proposed PhD thesis on his idea that stories have emotional shapes (Cumberg, 2010). Syuzhet.R and Sentimentr.R are two R libraries for ingesting plain text, segmenting it, creating a smoothed time series based upon the sentiment polarities of each segment, and visualizing these arcs.

SentimentArcs is an ensemble of three dozen sentiment analysis models from various families provided as a series of Jupyter notebooks in Python providing advanced functionality like customized smoothing, dynamic time warping, and automatic peak/valley detection and extracting text at key crux points.

2.3 Multimodal sentiment analysis

MSA is a research specialty that generalizes the popular traditional text-based sentiment analysis NLP task (Das and Singh, 2023; Lai, 2023; Qin, 2024). For example, the 5th annual MuSE 2024 MSA Challenge's cross-cultural humor detection task integrates video, audio and text transcripts to predict humor using multimodal features like facial expressions, body language, tone of voice, and verbal content (Amiriparian et al., 2024). However, text-centric LLMs are still arguably the most fundamental modality in MSA, while other modalities may contribute additional information with a lower signal/noise input (Yang, 2024).

The proliferation of both specialized and multimodal LMMs adding image, sound, and even video to core LLMs by OpenAI, Google and Meta are unifying MSA under a single model and API. However, the most performant MSA is still highly technical and interdisciplinary, integrating fields like natural language processing, AI, signal processing and neuroscience (Dai et al., 2017). For example, the winner of the most difficult MuSe 2023 task, MuSE-mimic (assessing the extent of approval, disappointment, and uncertainty in videos), achieved first place with a combination of traditional feature extraction, model architecture, and hyperparameter optimization that includes feature fusion at the levels of representations and model layers and across complimentary model outputs (Yi, 2023).

In carefully controlled situations like clinical studies or e-learning (Daza et al., 2023), multimodal integration can strongly reinforce the signal: goals are relatively straightforward, modalities align and affect signals are unbiased (e.g., physiological inputs like heart rate, skin conductance, and electroencephalogram readings). In many real-world situations, sentiment analysis of different modalities can often be noisy (Mao, 2022; Wu, 2024) or contradictory as is the case in irony, humor or sarcasm (e.g., the phrase "I love that" accompanied by an eyeroll and flat voice). Narrative and film add artistic elements such as an actor's interpretation of a character and directorial or cinematographic choices or editing room cuts that create dramatic suspense or tension through conflicting thoughts, emotions and behaviors expressed via different modalities.

Multimodal Emotion Recognition (MER) is a related term that uses verbal, physiological signals, facial, body gesture, and speech to distinguish between emotions, feelings, sentiments, and moods in both artistic and non-verbal ways (Kalateh et al., 2024). The AWESOME Research Project combines psychology, ML, multimedia, human computer interaction (HCI), and design to analyze the viewer's emotional response in real time using subtitles, audio and video modalities (Chambel, 2023).

2.4 Film narratives

Narrative and film theory explores the underlying structure and meaning of stories across various media. Narrative elements include plot (the causal sequence of events), characters, setting, themes (underlying messages), conflict, dialog, symbols, and more. When smoothed, the sentiment time series can indirectly reveal key points or cruxes of a story: peaks/valleys, rising/falling trends, abrupt changes, or other features. "The Shape of Stories" illustrates how diachronic sentiment analysis helps inform traditional literary analysis (Elkins, 2022).

Films use different techniques to convey emotion to viewers. Some of these techniques are more direct like dialog, speaking style, facial expression and body language. Others are more subtle like lightning, camera angle, setting and film edits. It is especially important to be aware of these elements and how the filmmaker, cast, cinematographer, set designer, costume maker, etc. are trying to guide viewers to specific feelings using these techniques. Just as there are specific cultural cues that invoke meaning through pragmatics, so too can film techniques signal unspoken and unseen meaning in a scene or still.

As challenging as language can be with latent meaning and context, the visual medium of film can add more degrees of artistic freedom to muddle meaning. For example, dialog can often be in dramatic opposition with elements like a character's internal beliefs, values, or physical situation. Postmodernism has influenced film since the late 20th century, breaking structure and conventions with fragmented, non-linear storytelling, the mixing genres, the breaking of the 4th wall to step out the narrative, and the subversion of established tropes and audience expectations.

Another challenge is symbolism: the cultural psychology of weather-related symbols like monsoon/rain carries different emotional descriptions in the UK vs. S.E. Asia (Shweder et al., 2008) while similar emotional scenarios can result in different facial expressions between Japanese and Euro-Americans (Bächle, 2022).

2.5 Models

Although DNNs date from Rosenblatt's perceptron (Rosenblatt, 1958), it took decades of advances in hardware, data, and algorithms to realize performance beyond proof of concept in the lab. Since 2012, DNNs have revolutionized computer vision, speech recognition, and natural language processing, by learning hierarchical representations from auto-regressively training on vast amounts of data (LeCun et al., 2015). The self-attention heads in Transformers can more efficiently capture long-range dependencies and most SOTA LLMs have been based on variations in this architecture, for example, RoBERTa (Liu, 2019).

Scaling laws arose from research showing that larger models, bigger quality training datasets, and more compute used in training (Gadre et al., 2024) can still improve performance in LLMs (Hoffmann, 2022). This has led to an exponential rise in training costs approaching $200 M USD for the current version of Google Gemini (Stanford University, 2024). To democratize AI research and maintain access, both the academy and industry have created OSS alternatives for large datasets like the Pile (Gao et al., 2020), pre-trained models like BLOOM (Scao et al., 2022), and fine-tuned models like llama3 (Touvron, 2023) and Mistral 7B (Mistral, 2023). MultiSA uses two of the leading OSS models released by Meta (Llava-llama3 for image to text) and Microsoft (Phi3 for text to text).

3 Methodology

Although SOTA commercial LLMs are rapidly improving, smaller OSS models lag and are generally more performant when fine-tuned on specific tasks or modalities. To accommodate both SOTA commercial OSS models, two parallel pipelines are used to independently process both the video and subtitle text as shown in Figures 1, 2. After the sentiment polarity of both video and subtitle text streams are processed, both time series undergo post-processing transformations and visualization steps shown in Figure 2. Details of these steps are explained in the following sections.

3.1 Source code pipeline

The processing pipelines in Figures 1, 2 correspond to sequentially executing these Python programs in the /src subdirectory of the MultiSA repository:

3.1.1 Video processing programs

- **common_start_convert_filmdb_txt_to_csv.py** convert the human-friendly input file. /data/dataset_yt_plain_small.txt to output a machine-friendly film dataset file. /data/dataset_film_small_details.csv.
- **videos_step1_download_videos.py** reads dataset_film_small_details.csv to download full-length copies of films from YouTube using PyTube.
- **videos_step2_split_videos_to_scenes.py** uses PySceneDetect to split all downloaded videos into ~200–400 scenes depending upon a threshold parameter tunable for each film. This threshold value can be automated, but currently is discovered by exploratory data analysis ranging between 9 for film The Mob and 30 for the film Royal Wedding.
- **videos_step3_extract_keyframes_from_scenes.py** uses the video processing library OpenCV to extract a keyframe from the midpoint of each scene.
- **videos_step4_get_keyframe_descriptions.py** calls the customized llava-llama3 model via the local ollama API to get a text description of every keyframe movie still.
- **videos_step5_get_sentiments_from_keyframe_descriptions.py** evaluates the sentiment polarity between [−1.0 and + 1.0] for each keyframe text description, generating three sentiment time series using local models VADER, TextBlob and a customized llama3 model.
- **common_step6_normalize_and_plot.py** (video sentiment arcs; set global PLOT_TYPE = 'videos') read all video sentiment polarity time series and processes them by a. normalizing with z-scores, and b. global smoothing/high-frequency noise reduction with 10% SMA.

3.1.2 Transcript processing programs

- **common_start_convert_filmdb_txt_to_csv.py** described above (this only needs to be run once).
- **transcripts_step1_get_srt.py** parse dataset_film_small_details.csv and use youtube_transcript_api to download all film transcripts in srt format and save under filename format /data/transcripts/{genre}/{title}_{year}.json.
- **transcript_step2_combine_in_dataframes.py** convert transcripts from srt json format to Pandas' DataFrame csv format and add utterance time midpoint for plotting.
- **transcript_step3_get_sentiments_from_transcripts.py** use text sentiment analysis models to calculate the raw sentiment values for every transcript utterance using the 3 SA models.
- **common_step6_normalize_and_plot.py** (**transcript sentiment arcs**; set global PLOT_TYPE = 'transcripts') read all transcript sentiment polarity time series and processes them by a. normalizing with z-scores, b. global smoothing/high-frequency

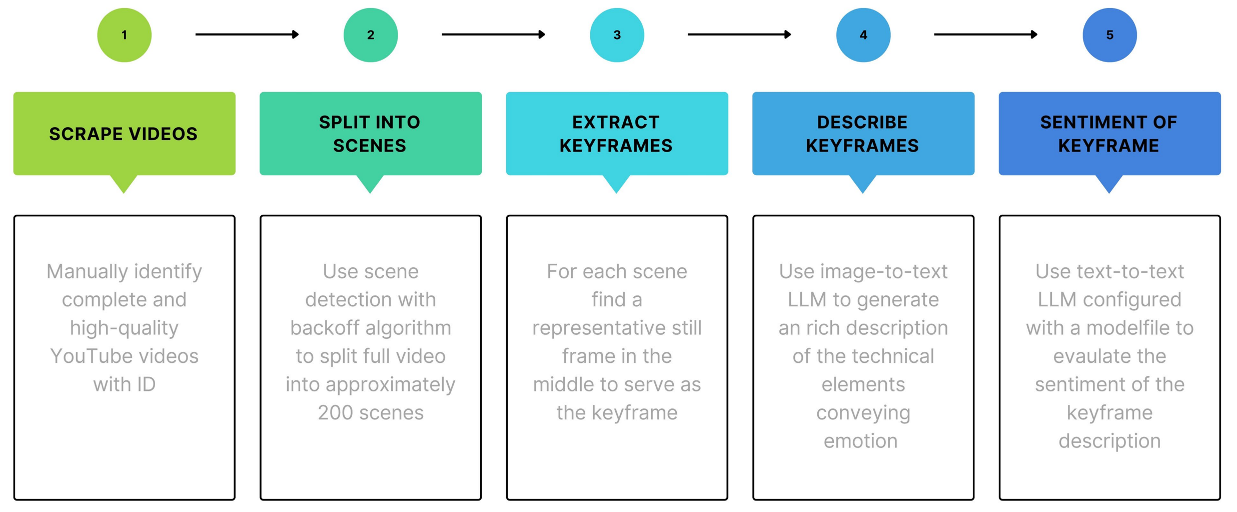
1
2
3
4
5
SCRAPE VIDEOS
SPLIT INTO SCENES
EXTRACT KEYFRAMES
DESCRIBE KEYFRAMES
SENTIMENT OF KEYFRAME
Manually identify complete and high-quality YouTube videos with ID
Use scene detection with backoff algorithm to split full video into approximately 200 scenes
For each scene find a representative still frame in the middle to serve as the keyframe
Use image-to-text LLM to generate an rich description of the technical elements conveying emotion
Use text-to-text LLM configured with a modelfile to evaulate the sentiment of the keyframe description

FIGURE 1
Video processing pipeline (keyframe images).

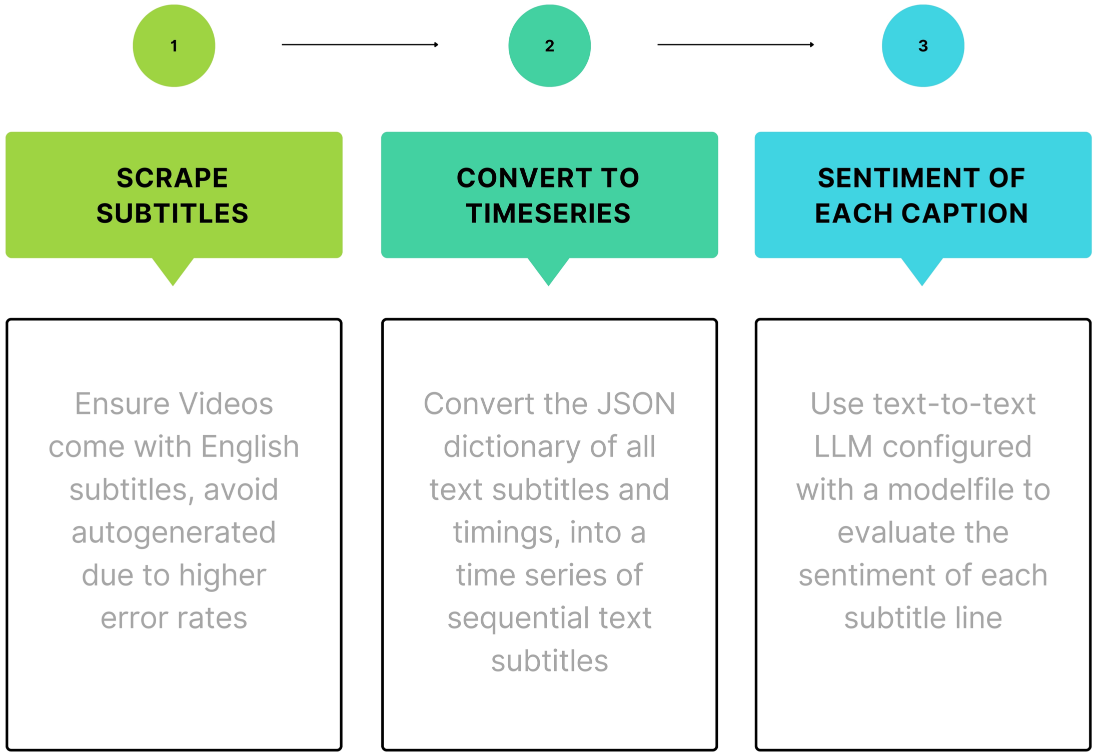
1
2
3
SCRAPE SUBTITLES
CONVERT TO TIMESERIES
SENTIMENT OF EACH CAPTION
Ensure Videos come with English subtitles, avoid autogenerated due to higher error rates
Convert the JSON dictionary of all text subtitles and timings, into a time series of sequential text subtitles
Use text-to-text LLM configured with a modelfile to evaluate the sentiment of each subtitle line

FIGURE 2
Transcript (text) processing pipeline (text).

noise reduction with 10% SMA, and c. local smoothing between unequally spaced datapoints for local smoothing/interpolations.

3.2 Data transformation sequence

As programs are executed in sequence and data is transformed along the pipeline, data exists in intermediate forms stored in the following subdirectories under /data:

3.2.1 Video data

- **/data/videos/{genre}/{title}_{year}/** contains mp4 video files of films downloaded by video_step1_download_videos.py sorted into subdirectories organized by the 8 genre names.
- **/data/keyframes/{genre}/{title}_{year}/** contains image files for the approximately 200–400 keyframes extracted from each scene from each movie sorted into subdirectories by genre then film title and year. Within the {title_year} subdirectory, all keyframe image files are sequenced by scene number (scene_no) using the filename convention **scene{no}_{title}_{year}.png** where leading and trailing numbers may be deleted to trim non-film frames.
- **/data/keyframes_descriptions/{genre}/{title_year}/** contains text files with image-to-text descriptions of each corresponding keyframe using the filename convention **scene{no}_{title}_{year}_description.txt.**
- **/data/keyframes_sentiment/{genre}/{title_year}/** subdirectories holds a Pandas' DataFrame export csv file with one row for each keyframe and columns for corresponding keyframe text descriptions and corresponding raw numerical text sentiment analysis values using VADER, TextBlob and Llama3.
- **/data/plots/{genre}** stores 3 files for every film's diachronic sentiment analysis using 2-step video sentiment analysis with keyframes and keyframe text descriptions:

(a) **{title}_{year}_normalized_sentiments.csv** stores a Pandas' DataFrame export csv file that contains with one row for each keyframe and columns for corresponding keyframe text descriptions with both raw and z-normalized text sentiment analysis values using VADER, TextBlob, and Llama3.
(b) **{title}_{year}_kde_plot.png** is a plot of the distributions of normalized sentiment values for the three text sentiment analysis models: VADER, TextBlob and Llama3.

- **{title}_{year}_sma10_plot.png** plots three diachronic sentiment analysis arcs over the sequence of keyframes using the two-step process of a. using an image-to-text to generate a text description of each keyframe, and then b. using NLP text sentiment analysis to generate a time series of sentiment analysis values for each keyframe.

3.2.2 Transcript data

- **/data/transcripts/{genre}/** stores the SubRip Subtitle (srt) text transcripts in json form for each film with filename format **{title}_{year}.json**.
- **/data/transcript_combined/{genre}** stores transformed srt transcripts as Pandas' DataFrame csv files with computed

utterance time midpoint with filename format **{title}_{year}_clean_transcript.csv**.

- **/data/transcript_sentiment/{genre}/{title}_{year}_clean/** directories each hold one exported Pandas' DataFrame csv file with a line for each utterance and corresponding time markers and 3 model raw sentiment values for text utterance with the filename format **{title}_{year}_clan_sentiment_transcript.csv**.
- **/data/transcripts_plots/{genre}** directories store 3 files for every film's diachronic sentiment analysis of each transcription utterance using textual sentiment analysis.

3.3 Hardware

All computing was done on a HP Victus 16.1″ gaming laptop with an Intel i7 CPU, 32GB RAM, and a low-power laptop version of NVIDIA GeForce RTX 4060 GPU with 8GB priced at $1,400 in July 2023. All experiments were run under the Windows Subsystem for Linux (WSL2) shell under Windows 11 Home, version 23H2 (Microsoft, 2023). Ubuntu 22.04.3 LTS was the Linux version installed due to its popularity, stability, and position as the default Linux installed under the Microsoft App Store. To accelerate both LLMs inference and video processing, NVIDIA CUDA 12.4 was installed (NVIDIA, 2024a), a compatible version of CUDA Toolkit (NVIDIA, 2024b), and then ffmpeg ver. 3.4.13 was recompiled with this CUDA support (NVIDIA, 2021).

3.4 Dataset construction

3.4.1 Overview

Unlike many other AI tasks like QA with labeled datasets and coding. SA generally lacks universal ground truth benchmarks beyond simple and narrow tasks like movie reviews and tweets (Lossio-Ventura, 2023; Stacchio et al., 2023). Consequently, the samples selected had to have as simple, clear, and interpretable sentiment arc as possible for HIL supervisors.

Post-WW2 Hollywood films are noted for their strong, simple, and almost predictable, narratives in comparison to more modern films influenced by postmodernist techniques like non-linear timelines and unreliable narration. The initial search parameters were for top grossing and/or top-ranked familiar Hollywood classics at the height of the golden era of Hollywood (1940–1955). This presented several difficulties as color was becoming widespread, genres were evolving, and many classic films from this era were usually not suitable for dataset inclusion (e.g., unavailable, low-quality, foreign language, no English subtitles, etc.).

Another problem with this initial search range was that film making was undergoing a rapid transition. Color films were becoming widespread, storylines and public sentiment grew more complex, and the popularity of specific genres was in flux. For example, film noir declined over this decade, but westerns/musicals were reaching their peaks. Thus, dataset construction focused on demonstrating the methodology of MultiSA using a few representative films rather than attempting to draw broad statistical generalities over a larger corpus by genre or year.

Several hundred top grossing films between 1940 and 1955 were initially considered. The first manual filter was to keep only films that could be primarily classified into one of eight genres including adventure, comedy, drama, film-noir, musical, psychological-thriller, romance, and western. Boundaries were fuzzy as many musicals had strong elements of romance or comedy and categories like dramas, romances, and adventures often subsumed others. The second and most selective cut was searching for high-quality, full-length videos available for download from YouTube. Ultimately, this resulted in a final clean dataset of 66 films: 8 adventures, 10 comedies, 5 dramas, 9 film-noir, 7 musicals, 9 psychological-thrillers, 9 romances, and 9 westerns listed in Supplementary Appendix A.

For example, here are three prominent 1951 Hollywood films selected to represent three different genres from among an initial pool of 66 films across eight genres screened. The final three genres range from the more formulaic Western (Rawhide) to slightly more variable plots found in Film Noir (The Mob). The musical genre is harder to characterize as it went through a rapid evolution from distinctly biopic performances like "Words and Music" (1948) to musical numbers written around more general narratives like "Singin' in the Rain" (1952). Our 1951 musical (Royal Wedding) tends toward the latter with songs both continuous with the narrative as well as discontinuous stories-within-stories performances. We chose popular classics to analyze due to their familiarity under the fair-use doctrine. We illustrate MultiSA using the 1951 movie Royal Wedding, which fell into the public domain due to failure to renew its copyright (PublicDomainMovies.com, 2023).

3.4.2 Video to still images

Initially, 80 popular Hollywood films from 1945 to 1955 were downloaded from YouTube using PyTube (2018). Fourteen were filtered out due to various quality issues such as inability to download completely, sound quality, subtitles, foreign language dubbing, poor video quality, etc. All accompanying subtitles were separately downloaded in srt-like JSON using youtube-transcript-api for greater flexibility (Depoix, 2018).

PySceneDetect was used to identify scene boundaries adjusting the threshold hyperparameter through an iterative backoff algorithm that could produce around 200–300 scenes per movie (SceneDetect, 2014). Finally, the same library was used to extract a still image from the midpoint of each scene as representative of the visual sentiment for that scene similar to SKFE (Gu et al., 2020). Each set of still images was visually inspected to remove opening or closing title cards or images not relevant to the narrative. Figure 3 illustrates four still keyframe images extracted from various scenes in the musical Royal Wedding.

3.5 Model and prompts

Two types and sizes of AI models are used in MultiSA. Small to midsize OSS AI models were used to evaluate text using Llama3 8B (Meta, 2023) and text plus video using Phi3 3B (Microsoft, 2024). These LLMs were installed locally using ollama v0.1.41 (Ollama, 2024) and accessed through the ollama's API Python library ollama-python (Ollama-Python, 2024). Diachronic video sentiment analysis is approximated by identifying scenes, extracting stills from the midpoint of all scenes as keyframes with associated timestamps, and then compiling this sequence of images with sentiment polarity into

FIGURE 3
Keyframe still images from Royal Wedding (1951) (Reproduced from 'Royal Wedding, 1951 Starring Fred Astaire and Jane Powell' - Public Domain Movies, licensed under CC0 Public Domain, https://publicdomainmovies.info/royal-wedding-1951-starring-fred-astaire-and-jane-powell/).

a time series of sentiment polarities as floating-point numbers between-1.0 (most negative) and 1.0 (most positive).

As a compliment and sanity check, sentiment analysis was also performed on each movie's textual dialog using simpler but popular sentiment libraries VADER using both lexicons and heuristic rules (Hutto and Gilber, 2014) as well as TextBlob using naïve bayes statistical ML (Loria, 2020).

Phi3 (text-to-text) was used as the LLM sentiment analysis engine and prompted to output a sentiment polarity (floating point in the range of [−1.0, 1.0]) based on input text. SFT and DPO were used to fine-tuned Phi3 for instruction following and safety (Microsoft, 2024) to produce chatty responses with introductions, explanations, definitions, summaries, conclusions, etc. This extraneous text was removed and Phi3 returned only the desired floating-point sentiment value by creating a customized model "Phi3sentiment" using ollama's modelfile feature (Ollama, 2023). The full modelfile is in the MultiSA github repository (Chun, 2023). Two key customizations are setting the hyperparameter "PARAMETER num_predict 5" to limit responses to five tokens in length and this "system" prompt:

> "You are a text sentiment analysis engine that responds with only one float number for the sentiment polarity of the input text. You only reply with one float number between-1.0 and 1.0 which represents the most negative to most positive sentiment polarity. Use 0.0 for perfectly neutral sentiment. You quietly but carefully think step by step to avoid extreme values with more insightful evaluations. Do not respond with any other text. Do not give a greeting, explanation, definition, introduction, overview or conclusion. Only reply with the float number representing the sentiment polarity of the input text."

Llava-llama3 (XTuner, 2024) was used as the image-to-text LMM to elicit a very structured, detailed, and evocative text description of keyframe images in every scene (~200 scenes/film). This detailed textual description was then fed into the existing Phi3 model to indirectly evaluate the sentiment polarity of each keyframe image. To consistently produce high-quality and relevant textual descriptions of each keyframe image, a custom ollama modelfile was used to create a customized model "llava-llama-3sentiment" based upon this "system" prompt in the Llava-llama3 modelfile:

> "You are an expert film maker and film critic specializing in film elements that convey emotion and sentiment including Facial Expression, Camera Angle, Lighting, Framing and Composition, Setting and Background, Color, Body Language and Gestures,

> Props and Costumes, Depth of Field, Character Positioning and Interaction, Visual Effects and Post-Processing".

Both SOTA GPT-4o (OpenAI, 2024) and larger OSS LMM (image to text) were initially used to directly evaluate the sentiment polarity of keyframe images. Given the many thousands of images analyzed when testing and filtering 66 full-length feature films, GPT4o API costs would be prohibitively expensive. The much smaller OSS LMM (image to text) models were too small to generate quality sentiment analysis floating point values directly from images.

Exploratory analysis showed smaller models lack the ability to generalize our narrow sentiment polarity task that is likely outside their fine-tuning dataset. On the other hand, much smaller OSS models have proven very capable when trained on narrow tasks, which we exploited with our two-step image sentiment analysis pipeline: (a) image to textual description (using Llava-llama3), then (b) text to sentiment polarity float values of the textual descriptions (using Phi3).

3.5.1 LLM text models

The Linux version of Ollama v.0.1.41 (Ollama, 2023) was the management framework for downloading and customizing all models. Models were considered based upon 3 criteria: size, performance, and popularity. Model size was constrained by our 8GB RTX 4060 to a max of about 7, 8, and 13GB LLMs with quantization. Performance was assessed with a combination of Hugginface.co Open LLM Leaderboard (Huggingface, 2024) and the LMSYS Chatbot Arena Leaderboard (Huggingface, 2024). Finally, popularity was determined by both Huggingface model card downloads as well as ollama.ai most popular model downloads.

We informally evaluated both leading free OSS and commercial SOTA LLMs. As of June 2024. This included OpenAI GPT4o (OpenAI, 2024), Claude Opus 3.0 (Anthropic, 2024a), Mistral 7B (Mistral, 2023), Phi3 (2024), and many fine-tuned, uncensored, or merged derivatives models like dolphin-mixtral (Hartford, 2024; Microsoft, 2024). To align findings with most AI benchmarks and research we chose to go with the extremely popular Llama3 8B Instruction fine-tuned model with 4-bit quantization (llama3:8b-instruct-q4_K_M, the default download for Llama3 on ollama.ai).

When prompted to evaluate a piece of text for sentiment polarity between-1.0 to +1.0 (most negative to most positive), the default response from llama3 often did not return the requested float value or embedded it within unwanted text like greetings, definitions, explanations, summaries, etc. The ollama .makefile configuration file and create command (GPU-Mart, 2024) was used to create a customized model (llama3sentiment) to overcome llama3's chat fine-tuning nature. A custom system prompt and hyperparameters constrained each response to be under five tokens representing a floating-point number polarity between-1.0 and 1.0.

3.5.2 LMM text and image models

GPT4o was the initial choice for the LLMs (text + vision) and in early tests it performed well. However, given the many thousands of images we processed over the course of this research and OpenAI's API cost model at the time, two free alternative OSS LMM were found to be surprisingly good at text descriptions of images. Of these two, vision enabled Llama3 variants, llava-phi3 (3.8GB) and llava-llama3 (5.5GB), the latter was chosen, llava-llama3:8b-v1.1-q4_0 downloaded from ollama.ai/library, due to greater parametric memory, slightly better performance and an acceptable inference speed.

The default download of the llava-llama3 model was extremely poor at evaluating the sentiment and returning a polarity between-1.0 and 1.0. In contrast with general SOTA commercial models perhaps 100x larger, smaller OSS models need narrow, well-defined tasks to excel. As such, image sentiment analysis was decomposed into two narrow, sequential tasks: (a). (image-to-text) describe the image in emotional and evocative terms using film-making techniques like facial expression, setting, and camera angle, and (b) (text-to-float) use standard NLP sentiment analysis on the textual descriptions of the image created in step (a) and return the sentiment polarity as a floating-point value within $[-1.0, 1.0]$.

An ollama modelfile was used to create a custom llava-llama3sentiment model focused on an image-to-text task of describing the image in the emotional language of various film-making techniques. Then llama3sentiment was used again for the second stage (b) to perform standard NLP sentiment analysis on this description.

3.6 Postprocessing time series

The sentiment values for each line of dialog in the captions and the textual description of each scene keyframe image represent one sample in a time series. For spoken dialog in transcripts, these samples are assigned a y-value based on sentiment and an x-value based upon the time that is the middle of the corresponding start and end times. Similarly, for keyframes taken from the middle of scenes, the y-value is the sentiment (image to text-description to floating point polarity) and the x-value is time based upon the time stamp of the image still extracted to create the keyframe. Figure 4 illustrates the steps for MultiSA simple post-processing of time series to generate cleaner data and interpretable visualizations.

Each sentiment time series is relatively noisy and must be smoothed. With no formulaic underlying generative mechanism, a non-parametric simple moving average (SMA) with a 10% sliding window was used to first filter out high-frequency noise and reveal the underlying sentiment story shape. Next, smoothed time series values are standardized using z-score normalization to center the mean = 0 and have a common sentiment scale to enable comparisons across different models and films. Finally, to deal with the unequal and irregular sampling intervals, univariate splines are used to curve fit between sparse data points. The transformations were done using the Python library SciPy ver. 1.13.1 (SciPy, 2020).

3.7 Similarity metrics

The agreement between different modalities in MSA can vary across applications from high coherence in tightly controlled clinical or physiological scenarios to low coherence in artistic depictions of narratives based on dramatic tension in film and television. Our MultiSA analyzes narratives using diachronic MSA and we propose two novel similarity/distance metrics to both measure the coherence/agreement between the sentiment arcs generated multiple models,

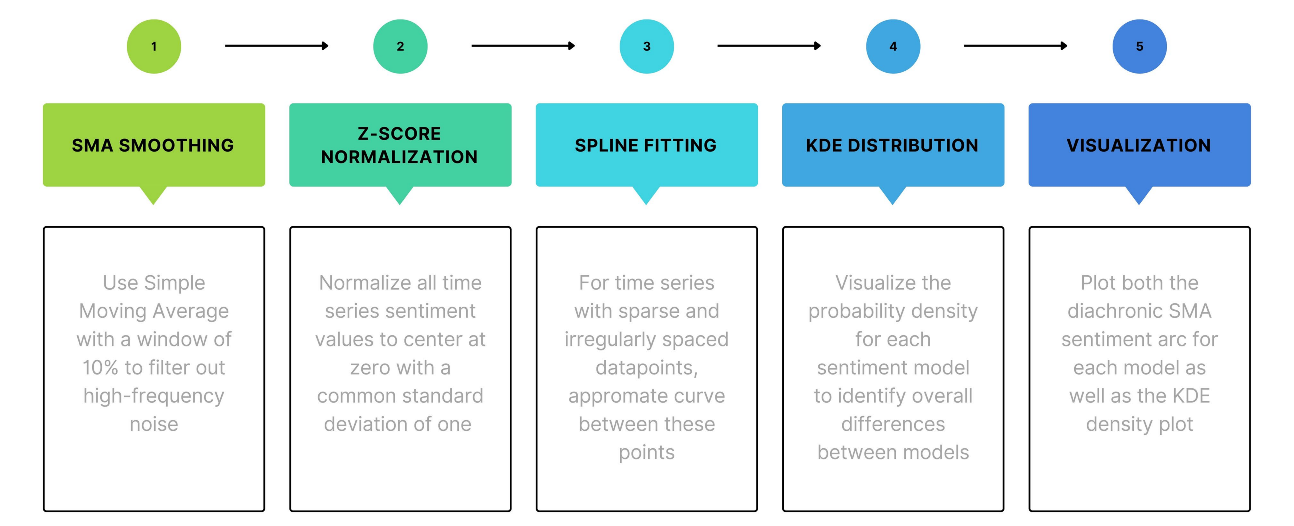
1
2
3
4
5
SMA SMOOTHING
Z-SCORE NORMALIZATION
SPLINE FITTING
KDE DISTRIBUTION
VISUALIZATION
Use Simple Moving Average with a window of 10% to filter out high-frequency noise
Normalize all time series sentiment values to center at zero with a common standard deviation of one
For time series with sparse and irregularly spaced datapoints, appromate curve between these points
Visualize the probability density for each sentiment model to identify overall differences between models
Plot both the diachronic SMA sentiment arc for each model as well as the KDE density plot

FIGURE 4
Post-processing pipeline for time series.

modalities and between OSS and SOTA Claude3 models. Our similarity metrics are used here to detect artistic discord among multimodal narrative sentiment arcs, but could also be used to detect, filter, or reweight noisy modalities for a more coherent aggregate signal in practical scientific applications.

Among dozens of generic time series distance metrics, we wanted one to capture the precise similarity anchored at exact times and one to measure more general shape similarity. Euclidian distance (Ed) and dynamic time warping distance (DTWd) respectively fulfilled these requirements. These both have the advantages of being relatively well known for reliable implementations. Ed has the advantages of being fast to compute based upon simple point-to-point distances while DTWd can transform unequal length time series and correct for temporal shifts and distortions.

In addition to these two fundamental metrics, we apply a series of more complex signal preprocessing transformations to enable more direct and meaningful comparisons including linear interpolations to dilate shorter time series, Winsorizing outliers, using Min-Max scaling, interpolating missing values, normalizing temporal distortions using DTW, non-parametric localized smoothing with LOWESS, and a final z-score normalization before plotting. See /src/analyze_step[1–3]* _coherence.py for the exact sequence of transformations applied to each metric to compute the three similarities: intermodal, intermodal, and open-claude3 benchmarks.

3.7.1 Euclidean distance

Euclidean distance is a standard metric for measuring the straight-line distance between two points in a multi-dimensional space. For time-series data, it is calculated by taking the square root of the sum of squared differences between corresponding points in two time-series.

Given two time-series $S_1 = \left[s_{1,1}, s_{1,2}, \ldots, s_{1,n}\right]$ and $S_2 = \left[s_{2,1}, s_{2,2}, \ldots, s_{2,n}\right]$, the Euclidean distance is defined as:

$$D_E\left(S1,S1\right) = \sqrt{\sum_{1}^{n}\left(s_{1,i} - s_{2,i}\right)^2}$$

To normalize the Euclidean distance, the raw distance is divided by the sum of the standard deviations of the two series:

$$Normalized\ \ Euclidean\ \ Distance = \frac{\sqrt{\sum_1^n\left(s_{1,i} - s_{2,i}\right)^2}}{\sigma\left(S_1\right) + \sigma\left(S_{2.}\right)}$$

where σ(S1) and σ(S2) are the standard deviations of $S_1 and\ S_2$, respectively.

The normalized Euclidean similarity, is then:

Euclidean Similarity = 1 – Normalized Euclidean Distance

which ranges from 0 (completely different) to 1 (identical).

3.7.2 Dynamic time warping distance

DTWd is a more flexible metric that can accommodate misaligned time-series data. It can account for temporal shifts and varying speeds in a time series. The DTW distance is computed by finding the optimal alignment between the time points of the two series that minimizes the cumulative distance.

For two time-series $S_1 and S_2$, the standard DTW distance is defined by constructing a cost matrix DDD where each element $D\left(i,j\right)$ represents the cumulative cost of aligning $S_{1,j} with\ S_{2,j}$

$$D\left(i,j\right) = cost\left(S_{1,i}, S_{2,j}\right) + \min\left(D\left(i-1, j-1\right), D\left(i-1, j\right), D\left(i, j-1\right)\right)$$

The optimal path through this cost matrix [from D(1, 1) to D(n,n)] gives the DTW distance.

To normalize the DTW distance, the raw DTW distance is divided by the maximum possible DTW distance, which is calculated between a constant series of ones and a constant series of zeros with the same length as the input series:

$$Normalized\ \ DTW\ \ Distance = \frac{DTWDistance\left(S_1, S_2\right)}{DTWDistance\left(1,0\right)}$$

The normalized DTW similarity is then:

DTW Similarity = 1 – Normalized DTW Distance

Like the Euclidean similarity, this metric also ranges from 0 (completely different) to 1 (identical). from 0 to 1, where higher values indicate greater similarity between the two time-series.

3.8 Visualizations

Figure 4 shows the post-processing pipeline used to generate the two main visualizations created for each of the 66 films saved in the /data/plots subdirectory. These were created using matplotlib ver. 3.9 (Matplotlib, 2012) and seaborn ver. 0.13.2 (Seaborn, 2012). The first is a kernel density estimation (KDE) plot showing the distribution of z-score normalized sentiment polarities across the 3 sentiment models for each film. Since these models are probabilistic in nature, the KDE model provides probabilistic distributions of sentiment polarity values from $n = 30$ samples to better visualize confidence and model agreement between VADER, TextBlob and Llama 3 subtitle sentiment analysis. The second is a 10% SMA smoothed plot for each model overlaying scatter plot points for each curve.

Each visualization plots all three sentiment analysis models: VADER, TextBlob, and LLM/LMM. Llama3sentiment is used for sentiment analysis on the text of transcripts. Llava-llama3sentiment is used to (1) describe keyframe images in text through the lens of specific film-techniques that convey emotion and (2) use Llama3sentiment to perform standard NLP sentiment analysis on this textual description. There are 132 KDE and 10% SMA plots for all 66 films saved under / data/plots/ organized under eight genre subdirectories.

3.8.1 Intermodel coherence

This gives visual qualitative and quantitative confirmation that our two-step OSS model pipeline, image-to-text description followed by text-to-text SA, is aligned with established baselines. We use two traditional SA models from different families, VADER (lexical and heuristic) and TextBlob (naïve bayes ML) to demonstrate our llama3 produces similar sentiment values for the given keyframe image description. These plots can be configured and generated by running /src/analyze_step1_video_coherence.py which outputs png files to / data/plots/{genre}/{title}_{year}_intermodel_coherence.png.

3.8.2 Intermodal coherence

Our two-step video sentiment time series gets video sentiment time series from /data/plots/{genre}/{title}_{year}_normalized_sentiments.csv while textual transcript sentiment time series are read from /data/transcripts_sentiments/{genre}/{title}_{year}_clean/{title}_{year}_clean_sentiment_transcripts.csv files. Both are inputs to /src/analyze_step2_video_transcript_coherence.py which creates a visual qualitative and quantitative metrics for the similarity between the video and textual transcription sentiment arcs in /data/plots/{genre}/{title}_{year}_video_transcript_coherence.png and save the normalized/smoothed mean video and transcript sentiment time series to /data/plots/{genre}/{title}_{year}_video_transcript_coherence.csv.

3.8.3 OSS and SOTA coherence

This measures the coherence between our two-step OSS video sentiment analysis and single-step video sentiment analysis using the SOTA Claude 3.0 model by Anthropic. The Claude3 sentiment analysis time series is generated using the Google Colab Jupyter notebook in /src/analyze_step3a_claude3_get_video_sentiments.ipynb. With a paid Anthropic API key, this script generates a Panda's DataFrame csv datafile {title}_{year}_claude-3-opus-20240229.csv for download which should then be moved into the /data/plots_claude3/ subdirectory. Configure and run /src/analyze_step3_video_claude3_coherence.py to read in both the 2-step open source and Claude3 video sentiment analysis time series, normalize them, compute their similarities, and save a png plot file to /data/plots/{genre}/{title}_{year}_coherence_claude3_open2step.png.

4 Results

4.1 Similarity metrics

The agreement between different modalities in MSA can vary across applications from high coherence in tightly controlled clinical or physiological scenarios to low coherence in artistic depictions of narrative in film and television. Our MultiSA analyses narratives using diachronic MSA and we use our two similarity/distance metrics to both measure the coherence/agreement between multiple models, modalities and between OSS and SOTA Claude3 models. Our similarity metric is used here to detect artistic discord among multimodal narrative sentiment arcs, but could also be used to detect, filter, or reweight noisy modalities in more practical scientific applications.

Among the dozens of time series distance metrics, we chose one to capture the precise similarity anchored at exact times and another to measure more general shape similarity. We applied approximately six transformations on Euclidian distance and Dynamic Time Warping distance fulfilled these requirements including using a non-linear sigmoidal transformation to compress our Ed and DTWd diachronic sentiment arc distance metrics between [0.0, 1.0]. The following three sections show how we used Ed, DTWd and traditional Pearson correlation to quantify similarities and coherence of diachronic sentiment arcs between different open SA models, modalities and open-*vs*-claude3 models.

4.2 Model agreement

One of the key findings is the strong level of agreement between the different sentiment arcs for all three films coming from entirely different genres (western, film noir, musical) as shown in Figure 5 (Rawhide), Figure 6 (The Mob), and Figure 7 (Royal Wedding). Since VADER, TextBlob, and LLMs all use very different methods (lexical plus heuristics, naïve bayes statistical ML, Transformers) this is a very strong indicator that there is a coherent latent sentiment arc these models are finding. If the models were less coherent, it would suggest we were fitting noise rather than a meaningful and consistent sentiment arc parallel to the narrative arc.

Among the three films, Rawhide has the most agreement between models while The Mob the least. This is explainable given the general genre and specific movies involved. As a film noir, The Mob has very muted sentiments and is filled with sarcasm, irony and deadpan acting. These make it much more difficult to accurately classify sentiments. In contrast, Rawhide follows a very strong typical Western story with high drama and emotion evident visually and in dialog. All

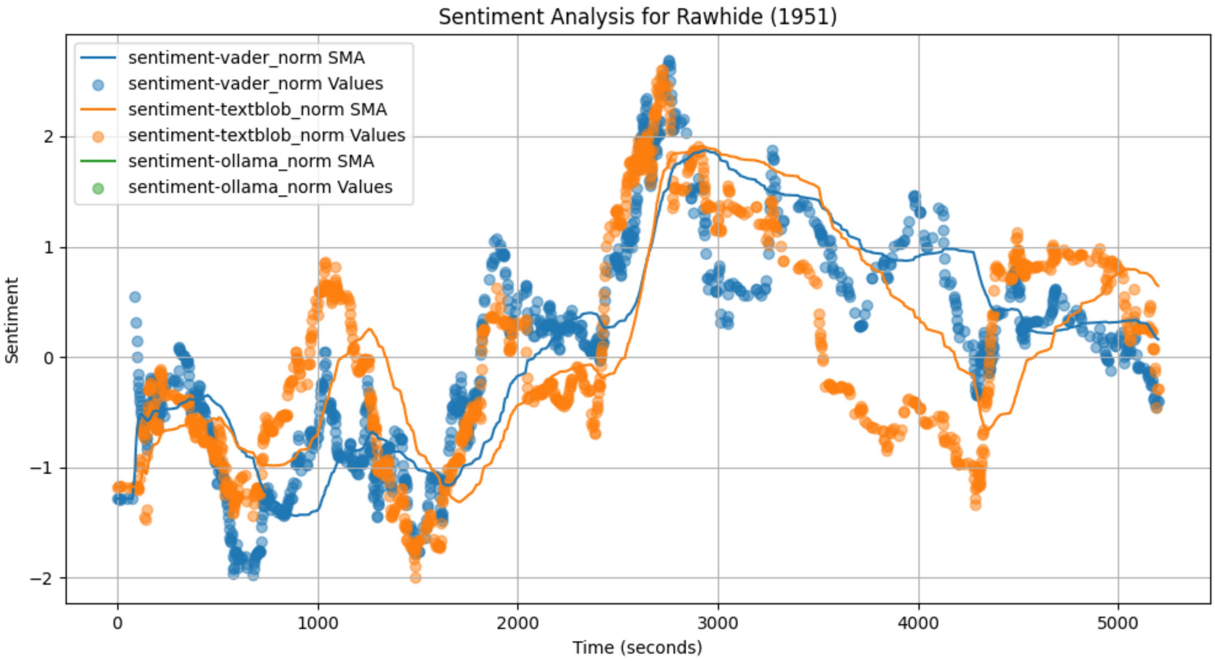

FIGURE 5
Transcript sentiment arcs (Western).

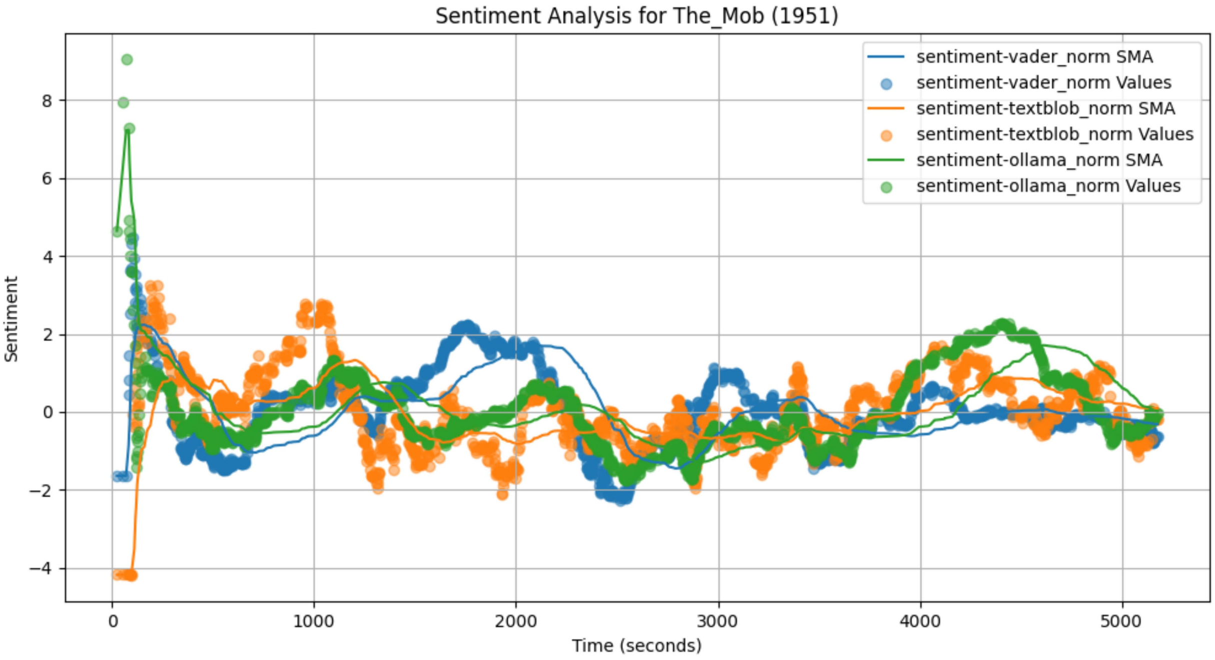

FIGURE 6
Transcript sentiment arcs (Film Noir).

three models should be able to more accurately pick up the same underlying sentiment arc resulting in greater coherence.

Figure 8 shows high intermodel coherence for diachronic sentiment analysis of the textual descriptions of keyframe images for the film Royal Wedding. This results in very high quantitative values across all 3-similarity metrics (Ed = 0.82, DTWd = 0.79 and Pearson's correlation = 0.94). The plot also qualitatively visually confirms the high degree of agreement between models. We use two traditional SA models from different families, VADER (lexical and heuristic) and TextBlob (naïve bayes ML) to demonstrate our llama3 produces similar sentiment values for the given keyframe image description. These plots can be configured and generated by running /src/analyze_step1_video_coherence.py which outputs .png files to /data/plots/{genre}/{title}_{year}_intermodel_coherence.png.

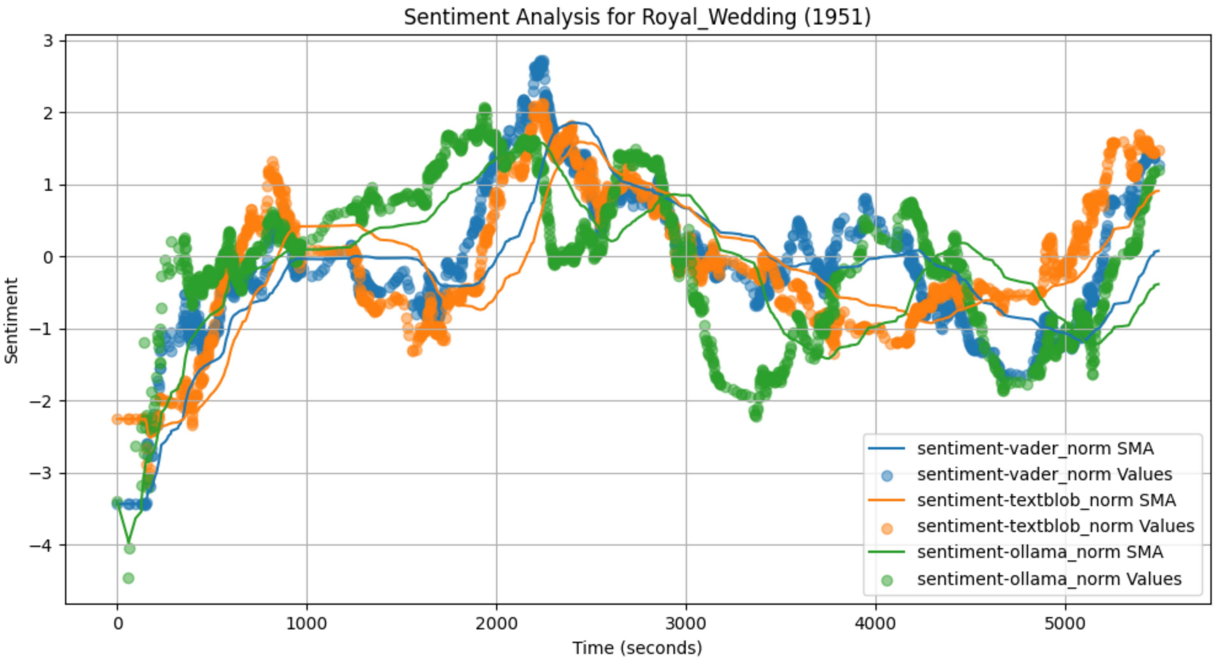

FIGURE 7
Transcript sentiment arcs (Musical).

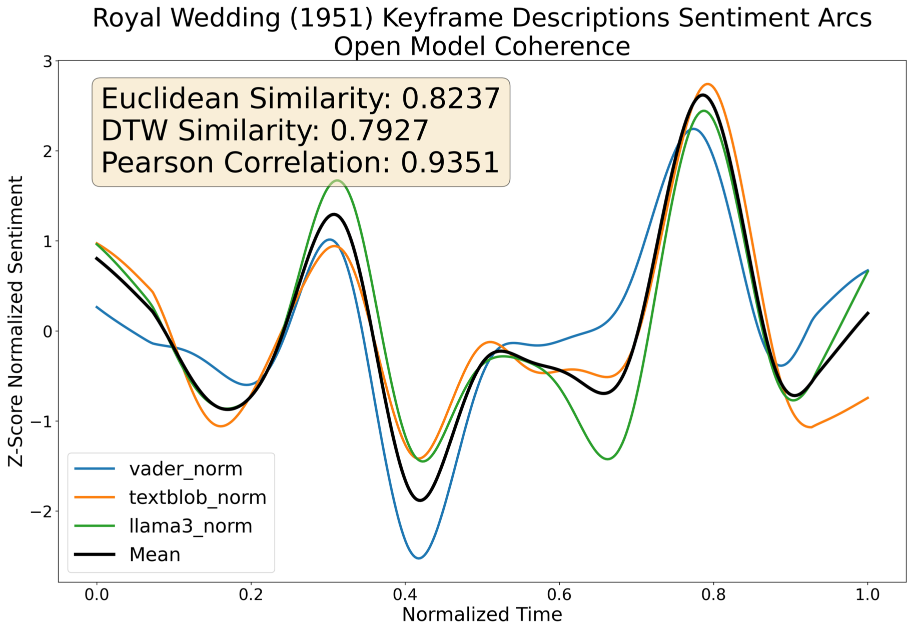

FIGURE 8
Multimodel sentiment coherence for keyframe descriptions.

4.3 Modality agreement

The agreement between different modalities in MSA can vary across applications from high coherence in tightly controlled clinical or physiological scenarios to low coherence in artistic depictions of narrative in film and television. Here we user our two novel similarity metrics, sigmoidal Euclidian and sigmoidal DTW, to quantify the dis/similarity between the video and text transcript sentiment time series.

As anticipated, the two sentiment arcs for the video and text transcript means have low agreements across all three metrics (Ed, DTWd and Pearson correlation) as shown in Figure 9. This has been seen in citied previous MSA research above, but also a natural result of different narrative modalities set in conflict with each other to

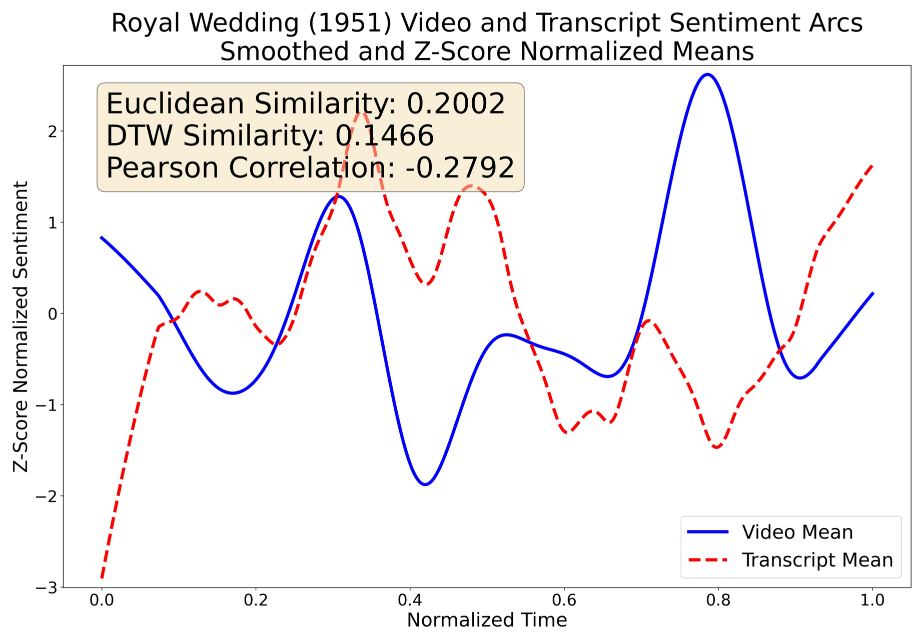

FIGURE 9
Multimodal video and transcript coherence.

create dramatic tension. For example, the surface plot may be captured in textual dialog, yet the countervailing subtext may be expressed either by actors (e.g., facial expression, body language, tone, etc) or the director and cinematographer (e.g., camera angle, lightening, color, editing, etc). The two biggest points of disagreement show the strength of each modality. The movie starts with rising action by taking their dance show to London for the royal wedding which is reflected in the text transcript while the video has sentiment plummeting. On the other hand, the video sentiment captures a peak corresponding to the colorful, non-verbal fantasy dance sequences before the end of the film which registered as a valley according to the transcript.

4.4 Open vs. SOTA model agreement

One of the major findings in this paper is that a sequence of two smaller (~7B each) specialized OSS models produce a very similar video sentiment arc compared to using a SOTA commercial LMM in Claude 3 Opus from Figure 10 plotting the video sentiment arc of the 1951 film Royal Wedding. Our similarity metrics (temporally precise Ed = 0.47 and more flexible shape similarity DTWd = 0.45) as well as the Pearson 0.45 show a moderate to high correlation. Qualitatively, it's important to note that both arcs track each other closely with sentiment falling at the beginning, a similar pattern of 3 peaks/4 valleys, and closing with rising action when lovers are suddenly reunited at the end.

4.5 Limitations

The MultiSA framework presented here only lays the core foundational components for a more interactive, explainable, and performant diachronic MSA system in general: a diverse and challenge dataset, customized quantitative metrics, intuitive qualitive visualizations, and a framework of extensible pipelines for multiple models and modalities. However, the interplay between multiple affective modalities and sources varies widely between fields and within applications which are not explored here. Even within film narratives, there is tremendous variation across genres, periods, filmmakers, and individual films themselves that remain to be mapped out and better understood. The carefully crafted disagreements between emotional modalities in film narrative point to the need to both augment MultiSA with more modalities (e.g., voice and music) as well as partition rich modalities like video (e.g., facial expression and body language).

Video scene boundary detection, segmentation, and keyframe extraction can be complex and the automated methods used here give only first approximation results. In addition, the 2-step OSS LMM to LLM pipeline may be dramatically improved by using models fine-tuned on specific tasks like analyzing film techniques and emotion with training datasets specific to Hollywood movies of the period, genre or even specific films. Many of the individual keyframes and descriptions are visibly incorrect, which large numbers and statistical smoothing is designed to filter out.

5 Future

This initial release of MultiSA currently focuses on video and image diachronic sentiment analysis and presenting novel modified metrics that can be used by human experts to analyze, explore, and critique the interesting interplay between MSA modalities. These same methodology of measuring the intermodal, intermodal and open-SOTA agreements can more broadly used to strengthen MSA model ensembles to identify dissonance, strategically involve human

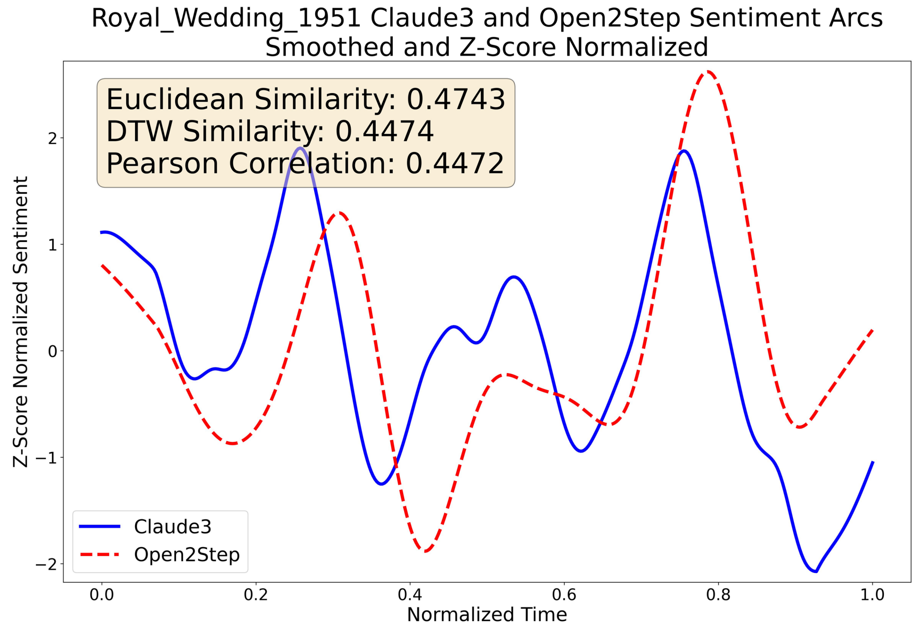

FIGURE 10
Two-step OSS vs. Claude 3 coherence (video).

experts, and reweight models to achieve higher performance in more clinical applications.

In addition to integrating these two modalities, our future research involves combining more narrative modalities like voice and musical soundtracks. Narrative subjects could be expanded to other source corpora beyond classic Hollywood movies, including conversations over social media (text/images/emojis), television speeches (video, text, voice) and medical narratives or customer support transcripts as we have already done many times with SentimentArcs using ensembles of dozens of textual, multilingual and visual SA models.

We also plan to expand upon video information by extracting information time series based on artistic and authorial cinematography choices like setting, color, lighting, and shot selection. Finally, we are exploring creating a far easier-to-use Docker image to avoid the complex and fragile setup of CUDA, CUDA Toolkit, ffmpeg recompilation, and customized ollama modelfiles that would present a barrier to most digital humanities scholars.

6 Ethics

MultiSA is a framework and benchmark to identify, extract, and analyze sentiment and narrative arcs in narratives, and like many technologies, could also be used by bad actors to discover ways to emotionally persuade, manipulate and deceive human via compelling narratives. MultiSA is part of the burgeoning field of affective AI research across disciplines and should be viewed within the lens of AI safety, human-AI alignment and AI regulation mentioned earlier.

There are no financial, commercial or other relationships involved in this research. SOTA Chatbots, including OpenAI's GPT4o and Anthropic 3.0 Opus were used to assist in developing, refining, and testing related code.

7 Conclusion

The paper illustrates a novel way and inexpensive way to use a two-step OSS model pipeline to extract video sentiment arcs from long-form narratives. We also introduce two novel modified similarity metrics to identify, analyze and potentially correct for incoherent modalities within any MSA application. We demonstrate these techniques using the very challenge, often contradictory narrative modalities of video and text transcripts across a dataset of 66 curated film classics and 8 genres from the golden age of Hollywood. We use highlight several films and genres to demonstrates SA processing pipelines with video, text transcripts, as well as across different SA models, modalities, and between OSS and SOTA commercial LMMs. Beyond the examples in this paper, the novel method presented here to detect diachronic multimodal sentiment arcs can generalize to any long-form narrative including political speeches, debates, long-trend topics in the news or on social media with varying degrees of MSA coherence. Practical applications could range from detecting latent persuasion techniques to improving emotional predictive models for better HCI communications. This also democratizes AI and can incorporate rapid improvements by using a flexible backend API built with free and OSS components.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://github.com/jon-chun/multisentimentarcs.

Author contributions

JC: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author declares that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The author thanks both the immediate providers of the OSS AI frameworks (ollama) and models (Microsoft Phi3 and Meta llama3) as well as all the contributors who have enabled the realization of this work (e.g., ffmpeg, llava-llama3, Huggingface.co, etc.). The still frames from *Royal Wedding* (1951) used in this paper are sourced from a film now in the public domain.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2024.1444549/full#supplementary-material

References

Abu-Nasser, B. (2017). 'Medical expert systems survey'. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082734 (Accessed on 28 September 2024).

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). *GPT-4 technical report.* arXiv preprint arXiv:2303.08774. Available at: https://arxiv.org/abs/2303.08774.

Afzal, S., Ali Khan, H., Jalil Piran, M., and Weon Lee, J. (2023). A comprehensive survey on affective computing: challenges, trends, applications, and future directions. *IEEE access* 12, 96150–96168. doi: 10.1109/ACCESS.2024.3422480

Agrawal, K. (2010). Study of the phenomenon of the Moravec's paradox. *arXiv preprint* arXiv:1012.3148.

Altarriba, J., and Kazanas, S.A. (2017). '*Emotions and expressions across cultures*'. *The International Encyclopedia of Intercultural Communication*, 1–10.

Amiriparian, S., Christ, Lukas, Kathan, Alexander, Gerczuk, Maurice, Müller, Niklas, Klug, Steffen, et al. (2024). 'The MuSe 2024 multimodal sentiment analysis challenge: social perception and humor Recognition'. Available at: https://arxiv.org/abs/2406.07753 (Accessed on 28 September 2024).

Anthropic, AI (2024a). 'Introducing the next generation of Claude (Claude 3.0)'. Available at: https://www.anthropic.com/news/claude-3-family (Accessed on 26 August 2024).

Anthropic, AI (2024b). 'Research'. Available at: https://www.anthropic.com/research#alignment (Accessed on 5 June 2024).

Ashwani, S., Hegde, Kshiteesh, Mannuru, Nishith Reddy, Jindal, Mayank, Sengar, Dushyant Singh, Kathala, Krishna Chaitanya Rao, et al. (2024). 'Cause and effect: can large language models truly understand causality? Available at: https://arxiv.org/abs/2402.18139 (Accessed on 28 September 2024).

Bächle, T. C. (2022). Faking it deeply and universally? Media forms and epistemologies of artificial faces and emotions in Japanese and euro-American contexts. *Convergence (Lond)* 29, 496–518. doi: 10.1177/13548565221122909

Birjali, M., Kasri, M., and Hssane, A. B. (2021). A comprehensive survey on sentiment analysis: approaches, challenges and trends. *Knowl.-Based Syst.* 226:107134. doi: 10.1016/j.knosys.2021.107134

Blanchard, E.G., Volfson, B., Hong, Y., and Lajoie, S.P. (2009). "Affective artificial intelligence in education: From detection to adaptation." *International Conference on Artificial Intelligence in Education*.

Brooks, J. A., Kim, L., Opara, M., Keltner, D., Fang, X., Monroy, M., et al. (2024). Deep learning reveals what facial expressions mean to people in different cultures. *iScience* 27:109175. doi: 10.1016/j.isci.2024.109175

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J. A., Horvitz, E., Kamar, E., et al. (2023). Sparks of artificial general intelligence: early experiments with GPT-4. *ArXiv* abs/2303.:12712. doi: 10.48550/arXiv.2303.12712

Burke, T. (2012). 'Commentary on Jonathan Haidt, the righteous mind | easily distracted', *Blogs.swarthmore.edu*, Available at: https://blogs.swarthmore.edu/burke/blog/2012/10/10/commentary-on-jonathan-haidt-the-righteous-mind/ (Accessed on 5 June 2024).

Chambel, T. (2023). "That's AWESOME: awareness while experiencing and surfing on movies through Emotions," *proceedings of the 2023 ACM international conference on interactive media experiences workshops*, pp. n.Pag.

Chun, J. (2021). 'SentimentArcs: a novel method for self-supervised sentiment analysis of time series shows SOTA transformers can struggle finding narrative Arcs', *ArXiv.org*, 18 October. Available at: https://arxiv.org/abs/2110.09454 (Accessed on 5 June 2024).

Chun, J. (2023). 'GitHub - Jon-Chun/Multisentimentarcs: a novel method to visualize multimodal AI sentiment arcs in long-form Narratives', *GitHub*. Available at: https://github.com/jon-chun/multisentimentarcs (Accessed on 25 August 2024).

Chun, J., and Elkins, K. (2023). eXplainable AI with GPT4 for story analysis and generation: a novel framework for diachronic sentiment analysis. *Int. J. Digit. Humanit.* 5, 507–532. doi: 10.1007/s42803-023-00069-8

Chung, F. (2024). '"20pc of Google": bizarre new site explodes', *News.com.au — Australia's leading news site*, 23 June. Available at: https://www.news.com.au/technology/online/internet/i-need-to-go-outside-young-people-extremely-addicted-as-characterai-explodes/news-story/5780991c61455c680f34b25d5847a341 (Accessed on 25 August 2024).

Cumberg, D. (2010). *Kurt Vonnegut on the shapes of stories.* [video] YouTube. Available at: https://www.youtube.com/watch?v=oP3c1h8v2ZQ (Accessed October 10, 2024).

Dai, Y., Wang, X., Zhang, P., and Zhang, W. (2017). Wearable biosensor network enabled multimodal daily-life emotion recognition employing reputation-driven imbalanced fuzzy classification. *Measurement* 109, 408–424. doi: 10.1016/j.measurement.2017.06.006

Das, R., and Singh, T. D. (2023). 'Multimodal sentiment analysis: a survey of methods, trends and Challenges. *ACM Comput. Surv.* 55, 1–38. doi: 10.1145/3586075

Daza, R., Gomez, L. F., Morales, A., Fierrez, J., Tolosana, R., Cobos, R., et al. (2023). MATT: multimodal attention level estimation for e-learning platforms. *ArXiv* abs/2301:09174. doi: 10.48550/arXiv.2301.09174

Depoix, J. (2018). 'Youtube-transcript-Api: this is a python API which allows you to get the transcripts/subtitles for a given YouTube Video', *PyPI*. Available at: https://pypi.org/project/youtube-transcript-api/ (Accessed on 25 August 2024).

Dylman, A., Champoux-Larsson, M. F., and Zakrisson, I. (2020). Culture, language and emotion. *Online readings in psychology and culture*, 4.

Ekman, P., Friesen, W. V., and Simons, R. C. (1985). Is the startle reaction an emotion? *J. Pers. Soc. Psychol.* 49:1416. doi: 10.1037/0022-3514.49.5.1416

Elkins, K. (2022). *The shapes of stories: Sentiment analysis for narrative*. Cambridge: Cambridge University Press (Elements in Digital Literary Studies).

EU Commission (2024). 'EU artificial intelligence (AI) act: first worldwide rules on AI', *EURAXESS*, 3 June. Available at: https://euraxess.ec.europa.eu/worldwide/lac/

news/eu-artificial-intelligence-ai-act-first-worldwide-rules-ai#:~:text=High-risk%20 AI%20systems%2C%20as%20well%20as%20certain%20users (Accessed on 5 June 2024).

Evans, N., and Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behav. Brain Sci.* 32, 429–448. doi: 10.1017/s0140525x0999094x

Fernández-Cruz, J., and Moreno-Ortiz, A. (2023). Tracking diachronic sentiment change of economic terms in times of crisis: connotative fluctuations of 'inflation' in the news discourse. *PLoS One* 18:688. doi: 10.1371/journal.pone.0287688

Gadre, S.Y., Smyrnis, Georgios, Shankar, Vaishaal, Gururangan, Suchin, Wortsman, Mitchell, Shao, Rulin, et al. (2024). 'Language models scale reliably with over-training and on downstream tasks', *ArXiv.org*, 13 March. Available at: https://arxiv.org/abs/2403.08540 (Accessed on 5 June 2024).

Gao, L., Biderman, Stella, Black, Sid, Golding, Laurence, Hoppe, Travis, Foster, Charles, et al. (2020). 'The pile: an 800GB dataset of diverse text for language Modeling', *ArXiv*, 31 December. Available at: https://arxiv.org/abs/2101.00027 (Accessed October 10, 2024).

GPU-Mart (2024). ' to customize LLM models with Ollama's Modelfile', *GPU servers Mart*. Available at: https://www.gpu-mart.com/blog/custom-llm-models-with-ollama-modelfile (Accessed on 26 August 2024).

Gu, X., Lu, L., Qiu, S., Zou, Q., and Yang, Z. (2020). Sentiment key frame extraction in user-generated micro-videos via low-rank and sparse representation. *Neurocomputing* 410, 441–453. doi: 10.1016/j.neucom.2020.05.026

Hartford, E. (2024). 'Cognitive computations (cognitive computations)'. Available at: https://huggingface.co/cognitivecomputations (Accessed on 26 August 2024).

Hoffmann, J. (2022). 'Training compute-optimal large language Models', ArXiv, 29 march. Available at: https://arxiv.org/abs/2203.15556 (Accessed October 10, 2024).

Hu, M., and Liu, B., (2004). "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04).* New York, NY, USA: Association for Computing Machinery, 168–177. Available at: https://doi.org/10.1145/1014052.1014073 (Accessed October 10, 2024).

Huggingface (2024). 'Open LLM Leaderboard - a hugging face space by open-Llm-Leaderboard', Available at: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard (Accessed October 10, 2024).

Hutto, C. J., and Gilber, E. (2014). VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proceed. Int. AAAI Conference on Web and Social Media* 8, 216–225. doi: 10.1609/icwsm.v8i1.14550

Jockers, M. (2019). '*Mjockers/Syuzhet*', GitHub, 11 May. Available at: https://github.com/mjockers/syuzhet (Accessed October 10, 2024).

Kalateh, S., Estrada-Jimenez, L. A., Nikghadam-Hojjati, S., and Barata, J. (2024). A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges. *IEEE Access* 12, 103976–104019. doi: 10.1109/ACCESS.2024.3430850

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Krugmann, J. O., and Hartmann, J. (2024). Sentiment analysis in the age of generative AI. *Departments of Labor, and Health, Education, and Welfare appropriations for 1961 Hearings before the Subcommittee of the Committee on Appropriations, House of Representatives, Eighty-sixth Congress, second session* 11:3. doi: 10.1007/s40547-024-00143-4

Lai, S. (2023). 'Multimodal sentiment analysis: a Survey', ArXiv, 12 may. Available at: https://doi.org/10.48550/arxiv.2305.07611 (Accessed on 4 January 2024).

Latif, S. (2022). 'AI-based emotion recognition: promise, peril, and prescriptions for prosocial Path', 14 November. Available at: https://arxiv.org/abs/2211.07290 (Accessed October 10, 2024).

Lavazza, A., and Inglese, S. (2023). The physiology of free will. *J. Physiol.* 601, 3977–3982. doi: 10.1113/jp284398

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., et al. (2023). Holistic evaluation of language models. *Ann. N.Y. Acad. Sci.* 1525, 140–146.

Liu, Y. (2019). 'RoBERTa: a robustly optimized BERT pretraining Approach', 26 July. Available at: https://arxiv.org/abs/1907.11692 (Accessed October 10, 2024).

Loria, S. (2020). 'Sloria/TextBlob', GitHub, 10 April. Available at: https://github.com/sloria/textblob (Accessed October 10, 2024).

Lossio-Ventura, J. A. (2023). A comparison of ChatGPT and fine-tuned open pre-trained transformers (OPT) against widely used sentiment analysis tools: sentiment analysis of COVID-19 survey data. *JMIR Mental Health* 11. doi: 10.2196/44437

Lynch, S. (2023). 'AI benchmarks hit saturation', *Stanford HAI*, 3 April. Available at: https://hai.stanford.edu/news/ai-benchmarks-hit-saturation (Accessed October 10, 2024).

Mabrouk, A. (2020). Deep learning-based sentiment classification: a comparative survey. *IEEE Access* 8, 85616–85638. doi: 10.1109/access.2020.2992013

Mao, H. (2022). Robust-MSA: understanding the impact of modality noise on multimodal sentiment analysis. *ArXiv* abs/2211:13484. doi: 10.48550/arXiv.2211.13484

Matplotlib (2012). 'Matplotlib: python plotting — matplotlib 3.1.1 documentation'. Available at: https://matplotlib.org/ (Accessed on 26 August 2024).

Meta (2023). 'Meta-Llama (Meta Llama 2)'. Available at: https://huggingface.co/meta-llama (Accessed on 25 August 2024).

Microsoft (2023). 'Set up a WSL development environment', Microsoft.com. Available at: https://learn.microsoft.com/en-us/windows/wsl/setup/environment (Accessed on 25 August 2024).

Microsoft (2024). 'Microsoft (Microsoft)', 22 August. Available at: https://huggingface.co/microsoft (Accessed on 25 August 2024).

Minaee, S. (2024). Large language models: a survey. *ArXiv* abs/2402:06196. doi: 10.48550/arXiv.2402.06196

Mistral, AI. (2023). 'Mistral 7B', Mistral.ai. Available at: https://mistral.ai/news/announcing-mistral-7b/ (Accessed October 10, 2024).

Momennejad, I. (2023). Evaluating cognitive maps and planning in large language models with CogEval. *ArXiv* abs/2309:15129. doi: 10.48550/arXiv.2309.15129

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2022). Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.* 56, 3005–3054. doi: 10.1007/s10462-022-10246-w

Murphy, H., and Criddle, C. (2023). Meta AI chief says large language models will not reach human intelligence. *Financial Times*, 8. Available at: https://www.ft.com/content/23fab126-f1d3-4add-a457-207a25730ad9 (Accessed October 10, 2024).

NIST (2023). 'U.S. artificial intelligence safety institute | NIST', NIST, 26 October. Available at: https://www.nist.gov/aisi (Accessed October 10, 2024).

NVIDIA (2021). 'Using FFmpeg with NVIDIA GPU hardware acceleration - NVIDIA docs', NVIDIA Docs. Available at: https://docs.nvidia.com/video-technologies/video-codec-sdk/11.1/ffmpeg-with-nvidia-gpu/index.html (Accessed: 25 August 2024).

NVIDIA (2024a). 'CUDA 12.1 release notes', Available at: https://docs.nvidia.com/cuda/cuda-toolkit-release-notes/index.html (Accessed: 26 August 2024).

NVIDIA (2024b). 'Installing CuDNN on windows — NVIDIA CuDNN V9.3.0 Documentation', Available at: https://docs.nvidia.com/deeplearning/cudnn/latest/installation/windows.html (Accessed: 25 August 2024).

Ollama (2023). 'Ollama/Docs/Modelfile.md at Main · Ollama/Ollama', GitHub. Available at: https://github.com/ollama/ollama/blob/main/docs/modelfile.md (Accessed: 25 August 2024).

Ollama (2024). GitHub, 29 February. Available at: https://github.com/ollama/ollama (Accessed: 5 June 2024).

Ollama-Python (2024). GitHub, 29 February. Available at: https://github.com/ollama/ollama-python (Accessed: 5 June 2024).

OpenAI (2024). 'Hello GPT-4o', OpenAI, 13 May. Available at: https://openai.com/index/hello-gpt-4o/ (Accessed: 25 August 2024).

Papers with Code (2024). 'Sentiment analysis', Paperswithcode.com. Available at: https://paperswithcode.com/task/sentiment-analysis (Accessed: 4 June 2024).

Patel, D. (2024). 'Francois Chollet, Mike Knoop - LLMs Won't Lead to AGI - $1,000,000 prize to find true solution', Dwarkesh Podcast. Available at: https://www.dwarkeshpatel.com/p/francois-chollet (Accessed: 25 August 2024).

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Plutchik, R. (1980). *A general psychoevolutionary theory of emotion*. New York: Academic Press, 3–33.

PublicDomainMovies.com. (2023). 'Royal Wedding, 1951 starring Fred Astaire and Jane Powell - public domain Movies', public domain movies, 18 September. Available at: https://publicdomainmovies.info/royal-wedding-1951-starring-fred-astaire-and-jane-powell/ (Accessed: 26 August 2024).

PyTube (2018). 'PyTube Github repo', Pytube.io. Available at: https://github.com/pytube/pytube (Accessed: 25 August 2024).

Qin, L. (2024). Large language models meet NLP: a survey. *ArXiv* abs/2405:12819. doi: 10.48550/arXiv.2405.12819

Quesque, F., Coutrot, A., Cox, S., de Souza, L. C., Baez, S., Cardona, J. F., et al. (2022). Does culture shape our understanding of others' thoughts and emotions? An investigation across 12 countries. *Neuropsychology* 36, 664–682. doi: 10.1037/neu0000817

Reagan, A. J. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.* 5, 1–12. doi: 10.1140/epjds/s13688-016-0093-1

Rinker, T. (2021). 'Cran/Sentimentr', *GitHub*, 12 October. Available at: https://github.com/cran/sentimentr (Accessed: 5 June 2024).

Romero, M. (2024). 'Mrm8488/Distilroberta-finetuned-financial-news-sentiment-analysis · hugging face', *Huggingface.co*, 21 January. Available at: https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis (Accessed October 10, 2024).

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and Organization in the Brain. *Psychol. Rev.* 65, 386–408. doi: 10.1037/h0042519

Salvi, F., Horta Ribeiro, M., Gallotti, R., and West, R. (2024). On the conversational persuasiveness of large language models: a randomized controlled trial. *ArXiv* abs/2403:14380. doi: 10.48550/arXiv.2403.14380

Saxena, A., Khanna, A., and Gupta, D. (2020). Emotion recognition and detection methods: a comprehensive survey. *J. Artificial Intell Syst*. 2, 53–79. doi: 10.33969/AIS.2020.21005

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., et al. (2022). BLOOM: A 176B-Parameter open-access multilingual language model. *ArXiv*, abs/2211.05100. Available at: https://arxiv.org/abs/2211.05100 (Accessed October 10, 2024).

SceneDetect (2014). 'Home - PySceneDetect', *Www.scenedetect.com*, 9 June. Available at: https://www.scenedetect.com/ (Accessed: 25 August 2024).

SciPy (2020). 'SciPy.org — SciPy.org'. Available at: https://scipy.org/ (Accessed: 26 August 2024).

Seaborn (2012). 'Seaborn: statistical data visualization — seaborn 0.9.0 documentation', Available at: https://seaborn.pydata.org/ (Accessed: 26 August 2024).

Shweder, R. A., Haidt, J., Horton, R., and Joseph, C. (2008). "The cultural psychology of the emotions: Ancient and renewed," in *Handbook of emotions* (3rd ed.). eds. M. Lewis, J. M. Haviland-Jones, and L. Feldman Barrett (New York: Guilford Press), 409–427.

Srivastava, A. (2022). 'Beyond the imitation game: quantifying and extrapolating the capabilities of language Models', *ArXiv*, 10 June. Available at: https://arxiv.org/abs/2206.04615 (Accessed October 10, 2024).

Stacchio, L., Scorolli, C., and Marfia, G. (2023). *Evaluating human aesthetic and emotional aspects of 3D generated content through eXtended reality*: *CREAI@AI*IA*.

Stanford University (2024). 'The AI Index Report – Artificial Intelligence Index', *Aiindex.stanford.edu*. Available at: https://aiindex.stanford.edu/report/ (Accessed October 10, 2024).

Strachan, J. W. A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., et al. (2024). Testing theory of mind in large language models and humans. *Nat. Hum. Behav.* 8, 1285–1295. doi: 10.1038/s41562-024-01882-z

Street, W. (2024). LLMs achieve adult human performance on higher-order theory of mind tasks. *ArXiv* abs/2405:18870. doi: 10.48550/arXiv.2405.18870

Sutton, R. (2019). *The bitter lesson*. Available at: http://www.incompleteideas.net/IncIdeas/BitterLesson.html (Accessed October 10, 2024).

Thaler, R. H. (2015). *Misbehaving: The Making of Behavioral Economics*. Available at: https://api.semanticscholar.org/CorpusID:152291962 (Accessed October 10, 2024).

Touvron, H. (2023). 'Llama: open and efficient foundation language Models', ArXiv, 27 February. Available at: https://arxiv.org/abs/2302.13971 (Accessed October 10, 2024).

Vaswani, A. (2017). 'Attention is all you need'. Available at: https://arxiv.org/abs/1706.03762 (Accessed October 10, 2024).

Wang, Y. (2022). 'A systematic review on affective computing: emotion models, databases, and recent Advances', ArXiv, 20 march. Available at: https://arxiv.org/abs/2203.06935 (Accessed October 10, 2024).

Wu, D. (2024). 'Resolving sentiment discrepancy for multimodal sentiment detection via semantics completion and Decomposition. *ArXiv* 2407:07026. doi: 10.48550/arXiv.2407.07026

XTuner (2024). 'Xtuner/Llava-Llama-3-8b-V1_1-Gguf · Hugging Face'. Available at: https://huggingface.co/xtuner/llava-llama-3-8b-v1_1-gguf (Accessed: 25 August 2024).

Yang, H. (2024). Large language models meet text-centric multimodal sentiment analysis: a survey. *ArXiv* abs/2406:08068. doi: 10.48550/arXiv.2406.08068

Yi, G. (2023). 'Exploring the power of cross-contextual large language model in mimic emotion prediction', **Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation**.

Zhang, W. (2023). 'Sentiment analysis in the era of large language models: a reality check'. ArXiv.org, 24 May. Available at: https://arxiv.org/abs/2305.15005 (Accessed October 10, 2024).

Zhang, Y., Yang, X., Xu, X., Gao, Z., Huang, Y., Mu, S., et al. (2024). Affective computing in the era of large language models: A survey from the NLP perspective. *ArXiv*, abs/2408.04638. Available at: https://arxiv.org/abs/2408.04638 (Accessed October 10, 2024).

Zhou, L., Gao, J., Li, D., and Shum, H. Y., (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Comput. Ling.* 46, 53–93.