

ARTICLE

Revisiting Weimar Film Reviewers' Sentiments: Integrating Lexicon-Based Sentiment Analysis with Large Language Models

Isadora Campregher Paiva¹, Josephine Diecke²¹ University of Amsterdam, ² University of Zurich

Keywords: Weimar cinema, sentiment analysis, ChatGPT, film reviews, LLM, NLP

<https://doi.org/10.22148/001c.118497>

Journal of Cultural AnalyticsVol. 9, Issue 4, 2024

Film reviews are an obvious area for the application of sentiment analysis, but while this is common in the field of computer science, it has been mostly absent in film studies. Film scholars have quite rightly been skeptical of such techniques due to their inability to grasp nuanced critical texts. Recent technological developments have, however, given us cause to re-evaluate the usefulness of automated sentiment analysis for historical film reviews. The release of ever more sophisticated Large Language Models (LLMs) has shown that their capacity to handle nuanced language could overcome some of the shortcomings of lexicon-based sentiment analysis. Applying it to historical film reviews seemed logical and promising to us. Some of our early optimism was misplaced: while LLMs, and in particular ChatGPT, proved indeed to be much more adept at dealing with nuanced language, they are also difficult to control and implement in a consistent and reproducible way – two things that lexicon-based sentiment analysis excels at. Given these contrasting sets of strengths and weaknesses, we propose an innovative solution which combines the two, and has more accurate results. In a two-step process, we first harness ChatGPT's more nuanced grasp of language to undertake a verbose sentiment analysis, in which the model is prompted to explain its judgment of the film reviews at length. We then apply a lexicon-based sentiment analysis (with Python's NLTK library and its VADER lexicon) to the result of ChatGPT's analysis, thus achieving systematic results. When applied to a corpus of 80 reviews of three canonical Weimar films (*Das Cabinet des Dr. Caligari*, *Metropolis* and *Nosferatu*), this approach successfully recognized the sentiments of 88.75% of reviews, a considerable improvement when compared to the accuracy rate of the direct application of VADER to the reviews (66.25%). These results are particularly impressive given that this corpus is especially challenging for automated sentiment analysis, with a prevalence of macabre themes, which can easily trigger falsely negative results, and a high number of mixed reviews. We believe this hybrid approach could prove useful for application in large corpora, for which close reading of all reviews would be humanly impossible.

Introduction

In the field of film studies, the majority of scholarly discourse revolves around a small number of canonized films. The reputation of such works can sometimes seem self-evident, particularly as the films become older than the living memory of scholars. Let us take as an example the Weimar classic

Das Cabinet des Dr. Caligari (1920). In the 2008 compendium *Weimar Cinema: An Essential Guide to Classic Films of the Era* Stefan Andriopoulos declares that “immediately after its release, *The Cabinet of Dr. Caligari* was acclaimed a masterpiece of German expressionist cinema” (13). This statement is seemingly so uncontroversial as to dispense the need for citation. Yet previous work by film historians who engaged with primary sources reveals a more contentious reception. In *German Expressionist Film*, John Barlow affirms that the film received considerable disapproval from those eager to assign fixed definitions of the cinematic arts in contrast to the theatrical (29). In *Dr Caligari at the Folies-Bergère*, Kristin Thompson cites a number of original sources in order to conclude that “[t]he French intellectual response to *Caligari* was mixed” (152). Later, she makes a broader statement about the film’s reception beyond this specific context, contending that “[s]uch views – for, against, and mixed – have characterized discussions of *Caligari* ever since” (154). Scholars focusing on the German and American critical response to the film found a rosier picture. Siegfried Kracauer states that German reviewers “were unanimous in praising *Caligari* as the first work of art on the screen” (71), while David Robinson finds that “[t]he German critics, almost without exception, ranged from favorable to ecstatic” (56), and “[t]he American press was by and large as enthusiastic as the German critics” (58).

These scholars have likely reached such conclusions based on close readings of primary sources, their divergence mostly caused by having access to different material, particularly when focusing on different countries. Yet, it is hard to know exactly what sources they accessed. As is typically the case for film historians, they cite only excerpts of, at most, a handful of reviews to make their points. Even if we can reasonably assume that they read many more reviews to reach their overarching conclusions, most of this material is left unacknowledged. That this means of reporting on qualitative research is the standard in historical work is understandable given that, for much of history, access to primary sources was mostly limited to what could be found in physical archives.

Though the digitization of archival materials is far from complete, sources like The Media History Digital Library (MHDL) have revolutionized archival research, not only by granting access to a treasure trove of digitized materials, but by making their texts fully searchable. This development sets the stage for more in-depth analysis of collections with computer assistance, as highlighted by Acland and Hoyt in *The Arclight Guidebook to Media History and the Digital Humanities*. We contend that, while film historians have embraced this newfound wealth of resources, our methodologies have not sufficiently adjusted to the new possibilities they confer. This article aims to serve as a proof of concept demonstrating the potential of digital methods for the analysis of historical film reception.

Firstly, since we maintain that digitalization means we can afford – and even have the responsibility – to be more transparent regarding our corpora, we are publishing our corpus and go into detail as to how it was built. While publishing corpora is common in quantitative analysis, we believe that qualitative research would also be better served by this, since sharing even a small corpus of the entirety of one's primary sources can make clear how many reviews one had access to and allow readers to potentially build on one's work or dispute one's conclusions.

Secondly, we believe that there is underutilized potential for “distant reading” (Moretti)¹ when it comes to film reception. In film studies, analysis of film reviews have been mostly limited to close reading. While this method yields much more nuanced and rich analyses, it comes at the expense of breadth of coverage. If we are only interested in the reception of canonized films, done one film and one country at a time, that is not a problem, but we believe there is much to be gained from going beyond that – not as an alternative methodology to supplant close reading, but to complement it. Following Denbo and Fraistat, we argue for a scalable approach, as “only by connecting the distant with the close can the potential of digital (...) analytics to address questions about culture be fully realized” (170). This approach is not confined solely to qualitative or quantitative methods, automated or manual processes; instead, it permits an iterative exploration involving both zooming in and zooming out of our data. Consequently, it avoids strict adherence to a single analytical level, allowing the examination of research objects and questions from diverse perspectives.

While the application of distant reading techniques to film reviews has been mostly absent in film historical work, that is not to say that the application of such methods to film reviews hasn't been done at all. Sentiment analysis of film reviews is a common topic in computer science papers (Maas et al.; Lu and Wu; Chen et al.). A particularly popular approach has been to use lexicon-based sentiment analysis. This is characterized by assigning texts an overall sentiment (typically positive or negative) through the use of pre-built dictionaries containing words and their associated sentiment scores. These techniques were developed especially to deal with vast amounts of consumer-generated data on the internet, such as product reviews. Film historians have seemingly – and, as we shall see, quite rightly – been skeptical of such techniques. The use of dictionaries which give a set sentiment value to a word regardless of the context in which it is used creates severe issues when it comes to nuanced critical texts, to the extent that the technique seemed to be ultimately useless for our purposes.

¹ While film studies has been undergoing a digital turn in the last 15 years, most of this has been dedicated to (moving) image analysis, such as average shot lengths (Tsivian and Gunars), film colors (Flueckiger and Halter) and even what some have termed “distant viewing” (Arnold and Tilton) or “distant watching” (Howanitz).

Recent technological developments have given us cause to re-evaluate the usefulness of automated sentiment analysis for historical film reviews. The release of ever more sophisticated Large Language Models (LLMs) – most famously of ChatGPT in November 2022 – has shown that their capacity to handle nuanced language could overcome some of the shortcomings of lexicon-based sentiment analysis. In light of the increasing use of LLMs in academic contexts (Bukar et al.; Hariri; Shen et al.; Sudirjo et al.), we were keen to assess the potential of ChatGPT and HuggingChat (an open source alternative) models for our research. The integration of LLMs into the domains of Digital Humanities and Computational Social Sciences introduces novel opportunities, capitalizing particularly on their apparent text comprehension capabilities alongside iterative inquiry methods (Ziems et al.). As we shall elaborate, some of our early optimism was misplaced: while LLMs, and in particular ChatGPT, proved indeed to be much more adept at dealing with nuanced language, they are also difficult to control and implement in a consistent and reproducible way—two things that lexicon-based sentiment analysis excels at.

Given these contrasting sets of strengths and weaknesses, we propose an innovative solution which combines the two, and has remarkably more accurate results. In a two-step process, we first harness ChatGPT's more nuanced grasp of language to undertake a verbose sentiment analysis, in which the model is prompted to explain its judgment of the film reviews at length. We then apply lexicon-based sentiment analysis with Python's NLTK library and its VADER (Valence Aware Dictionary for Sentiment Reasoning) lexicon to the result of ChatGPT's analysis, thus achieving systematic results.

The primary appeal of such an approach is typically seen to be its capacity for analysis of large corpora which go beyond what a human could feasibly read. Before one may use such an approach on a large corpus, however, it is necessary to test its reliability, comparing the automated results to those of careful manual annotation. The corpus for this test must be smaller and one which is intimately familiar to the authors. For this, we will use a corpus we prepared of 80 historical film reviews of three canonical Weimar films: *Das Cabinet des Dr. Caligari* (1920), *Nosferatu* (1922) and *Metropolis* (1927). In the next section, we will present this corpus, going over how it was selected and what it contains. Following that, we will explain and compare the different sentiment analysis methods we employed—encompassing ternary manual annotation (“positive”, “negative” and “mixed”), binary manual annotation (“positive” and “negative”) with LLM support, lexicon-based sentiment analysis with VADER and a hybrid model which integrates ChatGPT outputs with VADER sentiment analysis. We will show that these

approaches have variable accuracy² rates and ideal applications. We found that no purely automated approach reached perfect accuracy when compared to human annotation, though the two-step hybrid approach of ChatGPT-VADER showed a remarkable improvement when compared to lexicon-based sentiment analysis and could prove useful for large-scale application to film reviews.

Corpus Selection and Research Questions

Our first task was to locate film reviews pertaining to our chosen canonical Weimar films. To achieve this, we primarily utilized the online resources available through the MHDL and archive.org, as they offer a wealth of historical materials such as trade journals and fan magazines in English, German, French, and Spanish. We initially believed we would be able to gather the reviews for a large number of Weimar films automatically through web scraping, and thus create a corpus worthy of being called “big data”. This process was revealed to be far more complicated than initially anticipated. For starters, finding materials that actually relate to the films we were searching for is far from trivial. Film titles which are made up of common words, such as “Metropolis”, “Passion” (the English title for Ernst Lubitsch’s *Madame Dubarry*) or even letters, like “M”, make it difficult to identify the films without considerable manual work. Moreover, even if the word in question was indeed referring to the film we were interested in, instances of mere mentions, such as in lists of screenings, are a common occurrence, but offered limited value for sentiment analysis. Even harder would be to distinguish reviews from ads (by definition very positive) or plot summaries.³ It is also difficult to automatically recognize when a review starts and ends in a text—even for humans, the placement of text in a page feels somewhat arbitrary in many of the historical sources we are working with, with breaks in the columns at odd places and gaps of sometimes dozens of pages in the middle of a review. To make matters even worse, the quality of the Optical Character Recognition (OCR) of the texts is often subpar, affecting not only the quality of the digital text (which would make automated sentiment analysis of it problematic), but also undermining the reliability of the search itself. This OCR issue was particularly salient in relation to German texts obtained from the MHDL.

These challenges underscore the limitations of current technology when it comes to historical sources that were neither born-digital nor were digitized in a way that facilitates large-scale analysis. There is a reason why so much

² In the course of this paper, we will use the word “accuracy” to compare the results of the automated methods with those of our own manual annotations. Those are, of course, ultimately subjective judgments, and so the term must be understood in that context.

³ Even while reading the reviews carefully, this still proved difficult later. Indeed, as Eric Hoyt details in *Ink Stained Hollywood*, his history of Hollywood trade papers from 1915-1935, exhibitors were often rightly skeptical of the integrity of reviews in the trades. We ultimately decided to exclude only those items that made no attempts to disguise that they were ads and keeping all texts that followed a review format, that is, which provided at least a semblance of critical evaluation.

of the scholarly work done on film reviews has been done on born-digital material that is easily scrapable, like IMDb reviews (Shaukat et al.; Amulya et al.), rather than historical ones.

We subsequently altered our expectations and decided to create a corpus manually, and focus on fewer films. We picked *Das Cabinet des Dr. Caligari*, *Nosferatu* and *Metropolis* because those are some of the most canonized Weimar films according to several metrics.⁴ They have generated a considerable amount of film historical research that has – as seen in the introduction regarding *Caligari* – not led to a consensus about their reception, and we felt that openly publishing a collection of clean reviews which, though small when compared to big data, is still significantly larger than what most scholars have accessed to make their judgments, would still be a helpful contribution to the field.

We thus dedicated ourselves to a manual collection of reviews, eventually establishing a corpus of 80 reviews distributed as follows: 38 (47%) reviews for *Caligari*, 25 (31%) for *Metropolis*, and 17 (21%) for *Nosferatu* (Figure 1). While the majority of these (56%) were still obtained through the MHDL, other sources also proved important (Figure 2). A little-known website by the independent scholar Olaf Brill called filmbistoriker.de proved particularly helpful (contributing 32% of reviews) due to its provision of clean digital text of German-language Weimar-era film reviews split by movie. The third largest source of reviews were books, which reprinted historical reviews in full (Minden and Bachmann; Kaes et al.). Furthermore, our quest to broaden our collection of non-English language reviews led us to explore alternative sources, such as from our fourth online resource, filmportal.de (4%), maintained by the Deutsches Filminstitut & Filmmuseum (DFF). Unfortunately, filmportal holds very few reviews per film, and often only as an image with no OCR at all. While we applied OCR ourselves to those materials using the open-source software Tesseract, this was often futile when it came to poorly scanned and maintained sources, particularly those written in the German gothic script known as Fraktur.

Four languages are represented in our corpus: approximately the same number of reviews were written in English (34) and German (32), but there are also some in French (10) and Spanish (4) (Figure 3). Regarding the country of publication, a large number of reviews were published in Germany (33), followed by the USA (29), France (10), Spain (1) and Switzerland (1) (Figure 4). It is important to note that publication locations are not neatly aligned with languages. For instance, the magazine *Cine-Mundial*, which has four reviews in the dataset, presents a complex case of a Spanish-

⁴ For instance, these films constitute the top three most commonly assigned Weimar films in college syllabi, as derived from *OpenSyllabus* (Open Syllabus). Furthermore, they hold positions within the top 250 films featured in *Sight & Sound*'s esteemed "The 100 Greatest Films of All Time" list from 2012 (Sight and Sound).

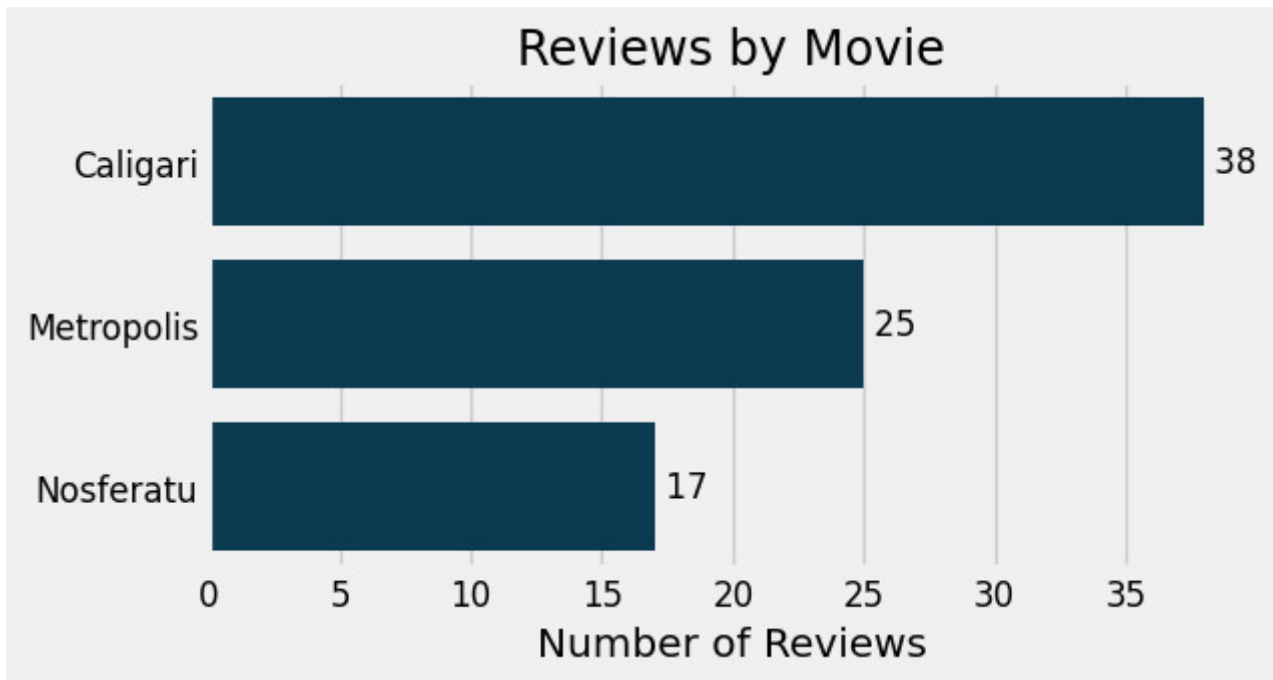


Figure 1. Distribution of Reviews by Film.

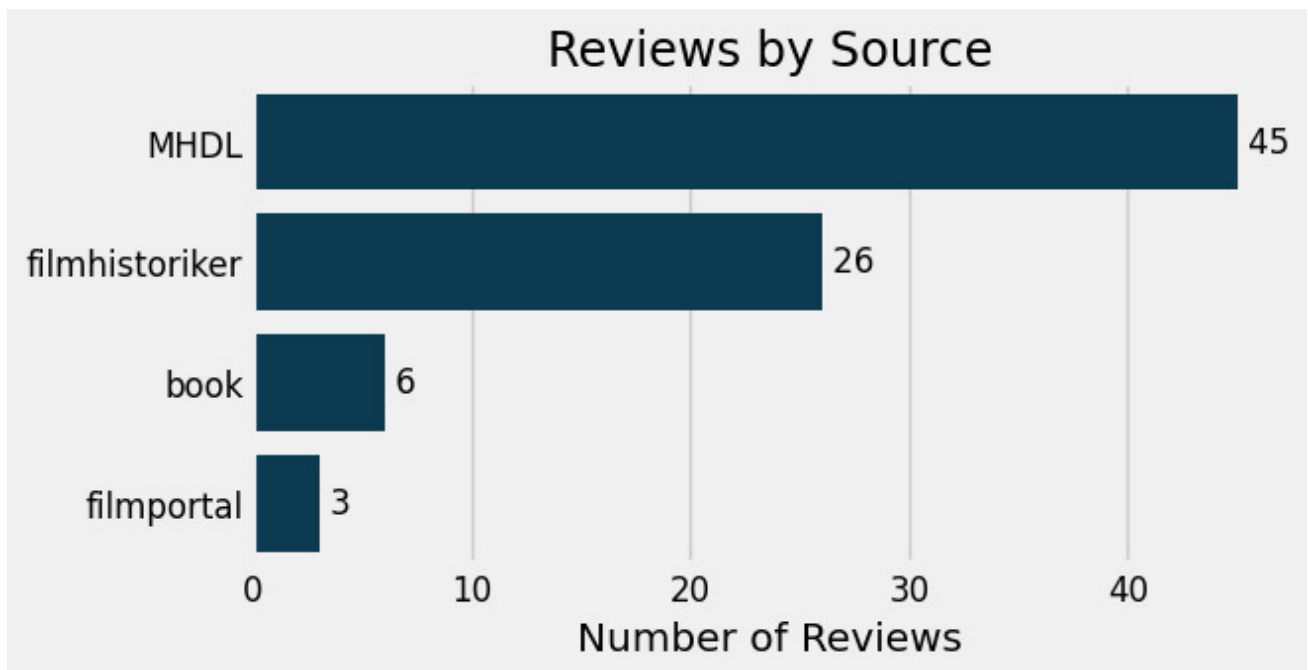


Figure 2. Distribution of Reviews by Source.

language publication targeted at the Spanish and Latin American market, but published in the USA. To ensure linguistic consistency for further processing, we translated all non-English language reviews to English. This was done with a combination of DeepL and manual checking of the results, as we are proficient in these four languages. While it would have been possible to use different models for different languages, we felt that building a uniform

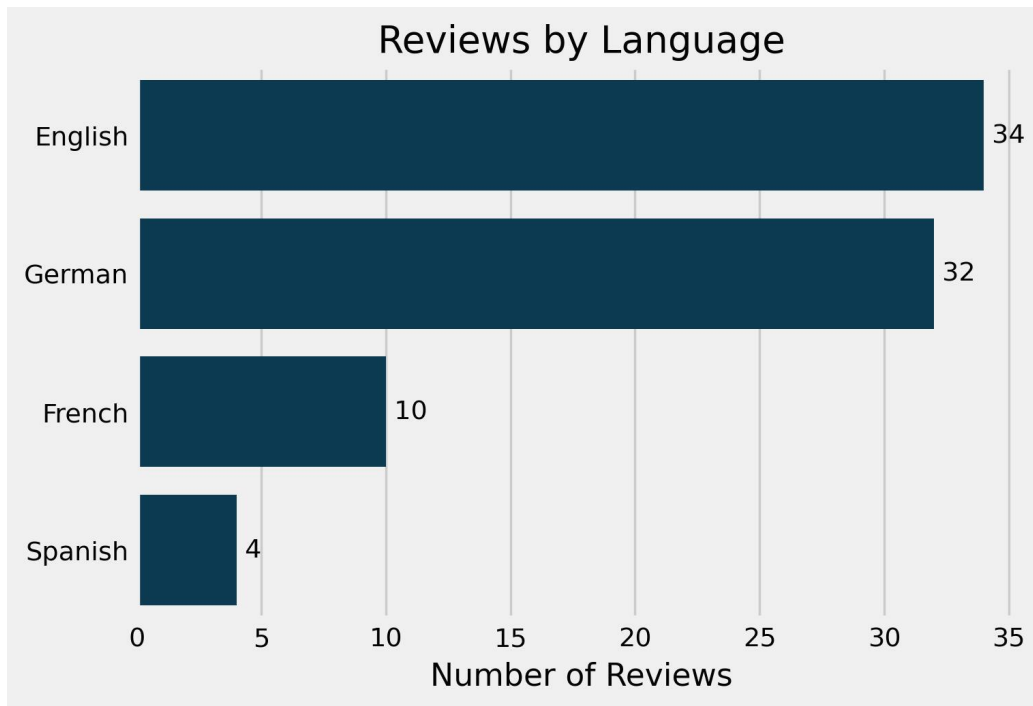


Figure 3. Reviews by Language.

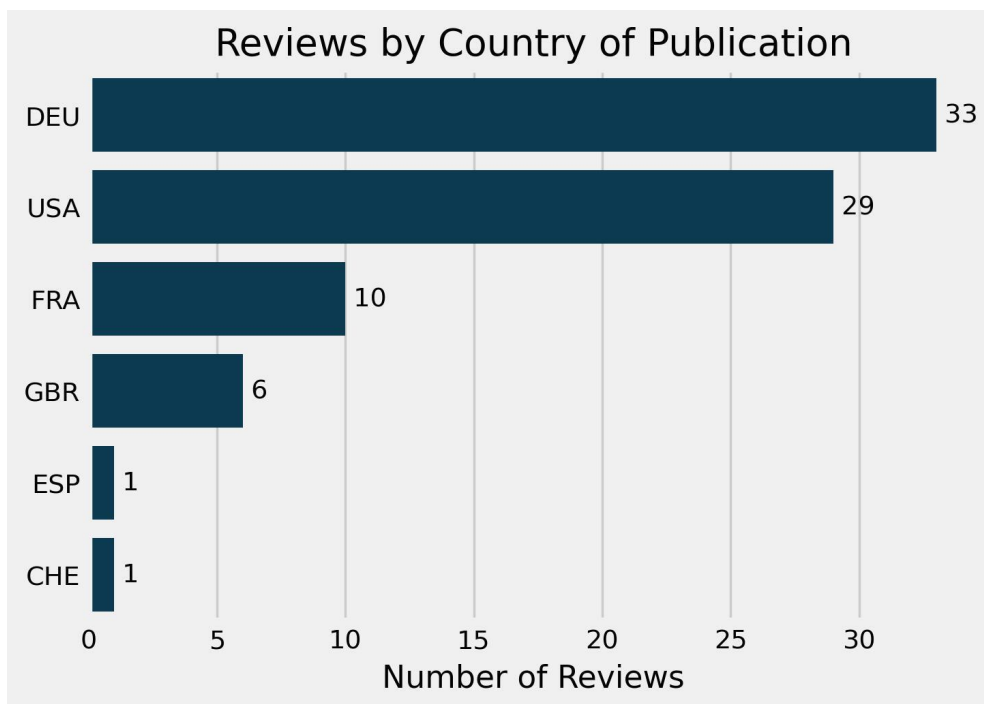


Figure 4. Reviews by Country.

corpus would be preferable to ensure consistency and comparability of the reviews. The comparison of sentiment analysis results between translated and non-translated reviews indicated that translation did not alter the outcome.

To manage the collected information, we established entries within the open-source reference management software Zotero. The subsequent steps within Zotero encompassed creating entries enriched with bibliographical details. We

diligently verified OCR content, which sometimes meant manually typing a review from scratch and eradicated duplicates, resulting in a refined collection of 80 reviews, each with accurate bibliographical information. For improved organization and analysis, we exported the Zotero data and used Open Refine for a final stage of data cleaning. Our final dataset, published here as an Excel table, was constructed by incorporating all the gathered data and metadata from the individual reviews. It includes metadata for the reviews, their full text in their original language and English translation, and the results of each of the sentiment analysis steps we took, which we will detail in the following sections.

Manual Annotation

Once we had a clean corpus, we conducted a manual sentiment analysis to establish a baseline for the succeeding computer-assisted phases. We deliberated over which sentiment categories were most useful and reflective of our objectives, and would afford the best basis for comparison with automated methods. The spectrum from positive to negative is the one typically used in sentiment analysis, so this seemed more fitting for comparison purposes than a star-rating, and also easier to assign. Inside that spectrum, we chose to include a mixed category, as it would allow for a recognition of reviews that are not easily categorizable in a binary, which are abundant in the corpus. We tried to characterize the review's overall sentiment, marking it as mixed only when there were substantial arguments made both in favor and against the picture. In other words, if a review was overwhelmingly positive or negative, minor comments to the contrary were not sufficient to get a "mixed" judgment.

Looking at the results of the manual sentiment analysis of our individual films ([Figure 5](#)), the following pattern emerges: *Caligari*'s reception was overwhelmingly positive, with 29 reviews (76.3%) marked as positive, 5 (13.2%) as negative and 4 (10.5%) as mixed. *Metropolis*, on the other hand, received mostly (52%) mixed reviews: 13 out of 25. This is followed by 7 positive reviews (28%) and 5 negative reviews (20%). Within the smaller corpus of reviews for *Nosferatu*, the reception was categorized as positive in 8 cases (47.1%), mixed in 6 cases (35.3%), and negative in 3 cases (17.6%).

Binary Manual Annotation with ChatGPT Support

Because of the high number of mixed reviews, we decided to add a separate column containing a binary classification (either positive or negative) of the reviews.⁵ Even mixed reviews typically lean slightly more to one side or the other, and having a binary rating in addition to the tripartite classification would allow us to better check the automated methods' accuracy. The

⁵ See column "Binary_Judgment".

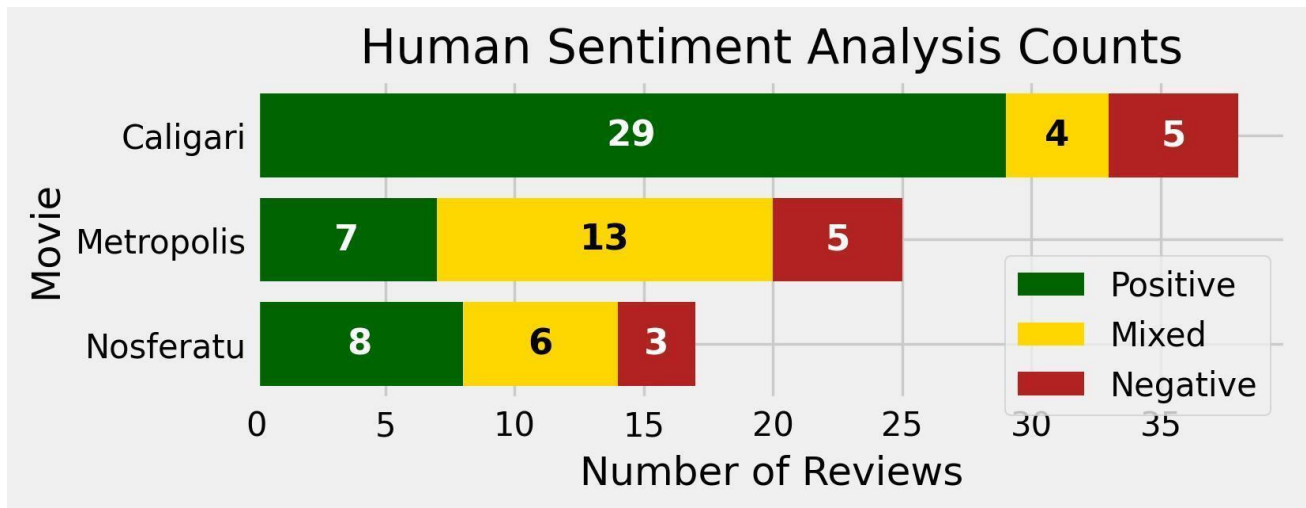


Figure 5. Sentiment Counts of Manual Annotation.

lexicon-based VADER model has a tendency towards polarization, so we wanted to be able to check when its classification was only slightly off the mark or very much so. For this binary annotation, we used ChatGPT in a manual and iterative manner to help us decide if a mixed review was *mostly* positive or *mostly* negative. The model thus served as an inter-rater reliability measure, in which one of the raters is not human.

This was the initial prompt used on all 23 “mixed” reviews: “The following is a film review. Would you say it is mostly positive or mostly negative regarding the film?”⁶ The answers typically contain around 130 words, and start with a sentence declaring “The film review appears to be mostly [positive/negative] regarding the film”, followed by statements justifying this judgment with direct quotes of the review’s relevant passages. This structure could also vary seemingly at random: in one case, ChatGPT’s answer came in the form of numbered bullet points (ID “TM6H4KLS”).

For the purposes of this binary manual annotation, we read through each of ChatGPT’s responses carefully, and we did not default to ChatGPT’s judgments, but rather used it as a separate rater and conversed with it when necessary to find an assessment we felt was fair. A compelling example of this arose in Herbert Ihering’s review “Der Metropolisfilm - Ufa-Palast am Zoo (Vorbericht)”, published on January 11, 1927, in the early edition of the Berliner Börsen-Courier (ID “TIRHYW52”). As indicated by the title, this is not a full-length review, but a Vorbericht (preliminary report) of Metropolis’ Berlin premiere. We felt that this distinction caused ChatGPT’s response to be slightly off the mark, so we followed up our initial prompt with an extra question:

⁶ ChatGPT’s answers are recorded in the column “ChatGPT Binary Answers”.

User:

The following is a film review. Would you say it is mostly positive or mostly negative regarding the film?

“A great film premiere. A lot of audience applause for Fritz Lang, the director, for Karl Freund, the operator, for the actors Alfred Abel, Heinrich George and Brigitte Helm. The film itself? A great effort with brilliant technical details for a banal, no longer viable idea. A future city and a stuffy past text.”

Chat GPT:

The film review appears to be mostly positive in its assessment of technical aspects and the people involved in the film, such as the director, operator, and actors. However, it also contains a negative note regarding the film's idea, describing it as “banal” and “no longer viable.” So, it's a mixed review with both positive and negative elements.

User:

The first portion of the review is actually referring to the premiere. The reviewer's opinion of the film itself is: “A great effort with brilliant technical details for a banal, no longer viable idea. A future city and a stuffy past text.” How would you characterize that, as mostly positive or mostly negative?

ChatGPT:

I apologize for the confusion earlier. Based on the review's assessment of the film itself, it is mostly negative. The reviewer mentions that the film has “brilliant technical details,” which is a positive aspect. However, they also describe the film's idea as “banal” and “no longer viable,” which are negative criticisms. So, overall, the review leans more towards the negative side.

ChatGPT's original answer is not exactly wrong, and indeed, a human might have easily made the same judgment. There is a reason, after all, why we had originally marked this review as “mixed”, as it is not a clear-cut case. We felt, however, that the reviewer's sentiment regarding the film seemed to differ from what the public at the premiere had demonstrated, and therefore pointed to the specific section of this short text that dealt with the film's quality as perceived by Ihering. The model's answer changed accordingly, matching our own perception.

Our judgment (and, after some prodding, ChatGPT's) of this Ihering's preliminary report seems to have been correct, since the author wrote a lengthier piece that day for the late edition of the same paper (ID

“TAKRH7XK”), where he makes his mostly negative opinion of the film clear. Indeed, this evening review was unanimously perceived as negative by all the sentiment analysis methods we employed.

Overall, we found that ChatGPT performed admirably at the task of helping us to assign binary sentiments to the reviews. ChatGPT’s output reliably summarizes the positively and negatively rated arguments in brief statements, often quoting relevant passages directly, and assigns a general sentiment based on the review’s overall tone. Because LLMs are designed to predict the most likely next word, they are particularly adept at producing text that represents what an average person might interpret as the overall sentiment of a review. For that reason, the model’s analysis is unlikely to be truly novel or especially insightful, but its very averageness makes it a useful tool as a sanity check for a human coder.

One disadvantage of this method is that it is not perfectly reproducible – were we to run the exact same prompt on our reviews again, ChatGPT’s responses would not be the exact same. Then again, neither would a human’s. While the exact wording would change, we did find that the overall sentiment judgments remained consistent in multiple runs of the same prompt and review combination.

LLM-only Sentiment Analysis

Given the excellent results we had with ChatGPT’s analysis of the reviews, one might think getting it to perform a one-word sentiment analysis would be trivially easy. If ChatGPT can deliver well-argued and nuanced analysis of a review’s overall sentiment, shouldn’t the labeling of reviews as “positive”, “mixed” or “negative” be a far simpler task? Unfortunately, that is not the case. ChatGPT and other LLMs are surprisingly resistant to following strict rules in their output and remaining consistent. We experimented with several different prompts trying to make the model behave in a dependable manner, to no avail. Changes could be minute – such as varying capitalization of the words and random addition of punctuation after a word – to severe – going off on a verbose analysis of the review such as the ones that had been useful in the previous section, but were not at all what we are hoping for at this stage.

Attempts to use open-source LLMs unfortunately had even worse results than ChatGPT. HuggingChat, a free and open-source LLM developed by HuggingFace, was not only unable to keep results concise and consistent in form (with similar issues regarding unexpected verbosity and inconsistent writing style), but the sentiments themselves proved to be inconsistent ([Table 1](#)). Changing the wording of the prompt hardly helped to make the results more formally consistent, and to make matters worse, the actual judgments were also erratic, as one can see by comparing the results with the same reviews and a slightly reworded prompt. In [Table 1](#), “prompt 1” was “The

Table 1. HuggingChat's inconsistent one-word sentiment analysis.

ID	prompt 1	prompt 2
2MBUUSRR	negative	Mixed.
L33Y2KDI	mixed	Negative
7PQP3M9K	Mixed.	positive
USR5PKR8	positive/mixed	positive

following is a film review: {text}. Assess the sentiment of this review in one word (positive, mixed, negative)". Prompt 2 was "The following is a film review: {text}. Assess the reviewer's judgment of the film in a single word (positive, mixed, negative)."

Lexicon-based Sentiment Analysis of Reviews with VADER

Confronted with such poor results, we decided to undertake a lexicon-based sentiment analysis of the reviews. Though we had low expectations for the quality of the assessments of the reviews with this method, we knew that at least its output would be consistent and reproducible, avoiding the pitfalls we had encountered attempting to generate consistent one-word sentiments with LLMs.

Rooted in the field of Natural Language Processing (NLP), lexicon-based sentiment analysis is a popular technique for extracting the emotional polarity of text, typically assigned either a score of -1 to +1, or words such as positive, negative, or neutral (Wankhade et al. et al.; Reborá). The method relies on predefined dictionaries of words and phrases that have associated sentiment scores. A text is thus analyzed as a collection of words, and the sentiment of the text as a whole is essentially an average of the individual words it contains. To be fair to this method, the way the words are arranged in the text does affect their value. The VADER model takes into account, for instance, if a word is in all caps (+ or - 0.733 of the word's base value), punctuation (for instance, + or - 0.292 to the sentence's score for each exclamation mark), and modifiers like negation, softening and contrast ([vaderSentiment.py](#)). For instance, in sentences containing the word "but", sentiment-bearing words before the "but" have their valence reduced by 50%, while those after the "but" increase their base values by 50% (ibid.).

These solutions are ultimately not enough to fully grasp the nuance of language. Lexicon-based models are known for their inability to capture contextual intricacies, irony, and sarcasm (Maynard and Greenwood; Farias and Rosso). The lexicon-based model struggles in particular with a corpus such as ours, which contains horror films like *Caligari* and *Nosferatu*. A word like "horrifying" is considered very negative by VADER (`vader_lexicon.txt`), with a valence of -2.7 (individual words are measured in a scale of -4 to +4). But in the context of a horror movie, that may well be a positive thing, as the film delivers the thrills it promises.

One may well argue that VADER was built to analyze Social Media Text (Hutto and Gilbert) and a different lexicon should be used to work on historical film reviews. A more tailored lexicon might slightly improve the quality of our results, but it would certainly not solve the overarching problem of a lexicon-based approach when applied to nuanced texts, however. What to do, for instance, with words like *horrific*? When attached to someone's acting or directing, it should rightly be judged as negative, after all. We ultimately chose VADER because it is the most widely used open source model for sentiment analysis due to its simplicity, speed, interpretability and lack of token limits (allowing longer texts). We could thus not only run it on our corpus with no fine-tuning, the same could be said for anyone who wanted to follow our model.

We applied VADER to our corpus of reviews, hoping to see just how accurate it could be when compared to our manual analysis.⁷ The resulting polarity scores range from -1 to +1.⁸ We also assigned sentiment values to those scores, in which values from -0.2 to +0.2 are mixed, while those below this band are negative, and above it are positive.⁹ We considered increasing the mixed interval in number because in practice, no scores were between these two numbers in this application, but decided against it because it did not improve the model's accuracy.

Given the lack of mixed cases in practice in the VADER sentiments, we measured the model's accuracy leniently: if the VADER sentiment matched *either* the original manual sentiment (negative, mixed or positive) *or* the manual binary sentiment (negative or positive), we considered its judgment to be accurate.¹⁰ We will go over the results of this method in depth when we compare it to our better performing hybrid model, but suffice it to say for now that the results of this direct application were unsatisfactory – VADER's direct sentiment analysis diverged from the human sentiment 27 times, that is, in 33.75 % of cases.

Hybrid Model: Lexicon-based Sentiment Analysis of Chat-GPT outputs

As previously discussed, we were impressed with the ability of LLMs, and particularly ChatGPT, to work with the nuanced language of historical film reviews and output well-argued analyses of even mixed reviews to indicate whether they leaned more towards positive or negative aspects. Yet, the raw power of LLMs proved difficult to harness in a consistent and reliable way for

⁷ See code in "sentiment_analysis.ipynb".

⁸ Results recorded in column "VADER_review_score".

⁹ See column "VADER_review_sentiment".

¹⁰ See code in "results_analysis.ipynb".

a one-word sentiment analysis of each review. The employment of a lexicon-based sentiment analysis had almost exactly the opposite characteristics: its results were clean, consistent and reproducible – they just happened to be wrong a third of the time.

Given these contrasting sets of strengths and weaknesses, we devised an innovative alternative approach, which combines the two methods for a fully automated sentiment analysis which has more accurate results, with an error rate of only 11.25%. We started from the realization that when we used ChatGPT to support manual analysis, it would typically output a relatively short text with declarative sentences which would accurately condense the sentiment of the reviews. The LLM-generated responses often include concise statements like “This film review appears to be [positive, negative, mixed],” and highlight the portions of the review which led it to come to this assessment, removing all extraneous information. We speculated that, when compared to the nuanced and complex language of the original reviews, the simplistic ChatGPT outputs would likely be much easier for lexicon-based sentiment analysis models like VADER to classify correctly. ChatGPT’s output is also more similar to what VADER was trained to work on – short texts in modern internet speech – compared to film reviews from the 1920s. Essentially, this would be an indirect sentiment analysis of the film reviews: rather than running the VADER model on the reviews directly, we would run it on the LLM-generated analysis of those reviews. This hybrid, two-step approach, would have most of the advantages of a lexicon-based sentiment analysis—systematic and fully automated—and would harness at least some of the advantages of LLMs’ more nuanced grasp of language.

Once more, we tried to use HuggingChat for this portion,¹¹ particularly because we would have to pay¹² to run this number of reviews through ChatGPT’s API, but found HuggingChat’s output to be subpar. The results to the prompt “The following is a film review. Would you characterize it as negative or positive? {review_text}”¹³ include responses like “I’m sorry, I am unable to understand your request. Can you please rephrase or provide more context?” (ID “DLNP7CTM”), which is why we abandoned this avenue.

Given these results, we again turned to ChatGPT. We did not use, as we had before, ChatGPT’s interface on the OpenAI website to run this portion of our experiment, utilizing its API instead to make this process fully automated through Python¹⁴. Since we wanted to maximize the accuracy of results and

11 See code “huggingChat_API_reviews.ipynb.”

12 For a small corpus like ours, this was ultimately not an issue, since running ChatGPT for *all* the tasks we used it for during the period of this paper only cost US\$0.32.

13 The results are recorded in the column “HuggingChat API Answers” of the Reviews Corpus table.

14 See code “GPT_API_all_reviews.ipynb.” To avoid cherry-picking, we ran this code only once, with the column “ChatGPT API Answers” of our corpus table containing the results.

avoid, to the best of our ability, that the sometimes dark subject-matter of the films in question would generate falsely “negative” results, we applied the ensuing prompt to all the reviews: “The following is a film review. Is it negative or positive? Explain your reasons, and keep in mind that negative emotions may be warranted depending on the genre of the film, and do not necessarily signify a negative review. Review: {text}”.¹⁵ We ran this request on only once, to avoid cherry-picking the best gene

We manually checked ChatGPT’s output and found that this prompt proved effective, producing well-reasoned summaries of each review’s overall sentiment. We subsequently ran the VADER model on the ChatGPT output, following exactly the same steps as what had been used directly on the film reviews.¹⁶ The analysis of the results will be discussed below, as part of a comparison and evaluation of all the methods we utilized.

Evaluation of Sentiment Analysis Methods

Comparing the results of the two automated sentiment analysis methods leaves little question as to which is more accurate. With a deviation rate of only 9 cases (11.25%) from the manual judgments, the VADER sentiment analysis of ChatGPT outputs proved to be considerably more reliable than the results of VADER when applied directly, which erred in 27 cases (33.75%). Most of VADER-only’s mistakes (17, or 62.9%) were due to incorrectly assigning a positive sentiment to a review that was originally mixed or negative ([Figure 6](#)). VADER-ChatGPT’s mistakes were much more balanced, with five false positives and four false negatives.

The direct and indirect VADER analyses yield a score ranging from -1 for negative to +1 for positive.¹⁷ Both methods had a tendency towards extremes. Though we assigned a mixed score between -0.2 and +0.2, in practice, mixed judgments were never observed in the VADER-only analysis and occurred only three times in the VADER-ChatGPT analysis. Interestingly, in all three of these cases, the manual judgment had indeed been categorized as mixed. In other words, not a single review was falsely categorized as mixed.

Examination of Interpretation Discrepancies per film

In total, there are 30 reviews out of 80 for which at least one of the automated sentiments diverge from the manually annotated sentiment (either binary or tertiary), that is, 37.5% of the time. In order to gain insight into the reasons behind the divergences, we will hone in on those 30 cases below. It is helpful to closely examine the deviations within individual films and reviews.

¹⁵ The resulting output was saved in the column “ChatGPT API Answers”.

¹⁶ See code “sentiment_analysis.ipynb”.

¹⁷ See columns “VADER_review_score” and “VADER_ChatGPT_score”.

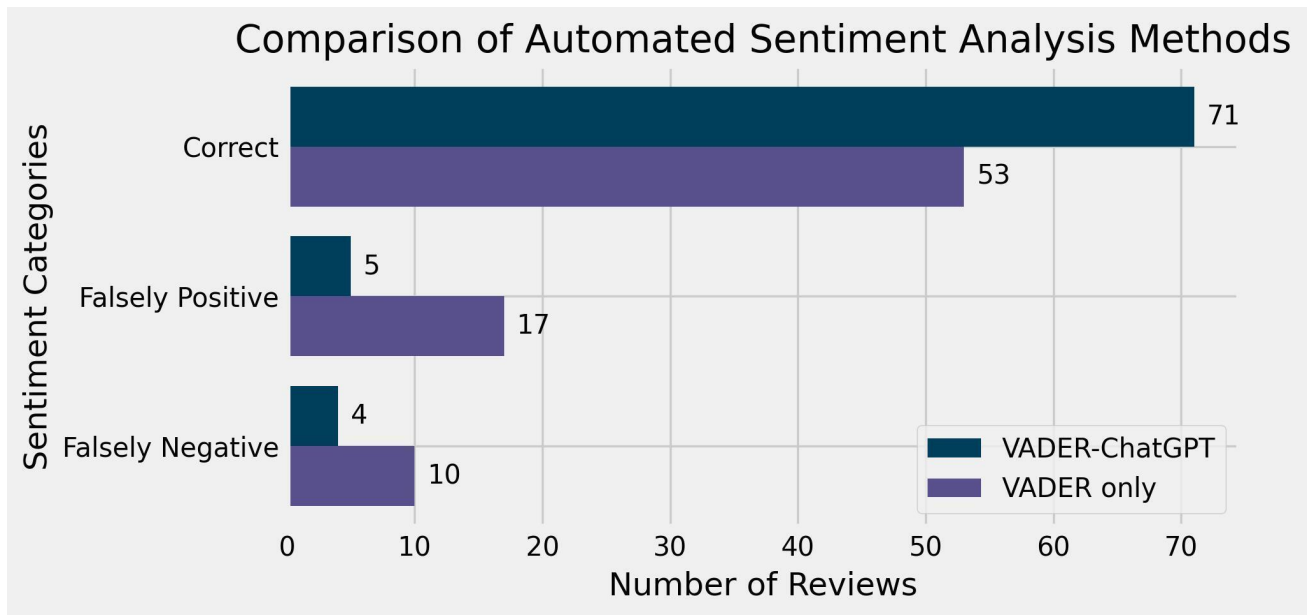


Figure 6. Comparison of Automated Methods' Accuracy.

ID	Movie	Manual_Judgment	Binary_Judgment	VADER_ChatGPT_sentiment	VADER_ChatGPT_score	VADER_review_sentiment	VADER_review_score	GPT_Error_Check	VADER_Error_Check
MXZKVQ6B	Caligari	negative	negative	negative	-0.610300	positive	0.997500	Correct	Falsely Positive
G2XQWE7E	Caligari	positive	positive	positive	0.984200	negative	-0.538700	Correct	Falsely Negative
LU2I7PCZ	Caligari	negative	negative	negative	-0.916900	positive	0.270900	Correct	Falsely Positive
PJSEZGQ4	Caligari	negative	negative	positive	0.971900	positive	0.452700	Falsely Positive	Falsely Positive
FKFS95J3	Caligari	mixed	positive	positive	0.956500	negative	-0.963100	Correct	Falsely Negative
WYMPWWPX	Caligari	negative	negative	negative	-0.694300	positive	0.889200	Correct	Falsely Positive
INTVKS32	Caligari	positive	positive	positive	0.917900	negative	-0.860600	Correct	Falsely Negative
VNXHI9SG	Caligari	positive	positive	negative	-0.361200	negative	-0.986100	Falsely Negative	Falsely Negative
JDMUZ7Z7	Caligari	positive	positive	positive	0.970500	negative	-0.579300	Correct	Falsely Negative
FWT8J86D	Caligari	positive	positive	positive	0.535800	negative	-0.995500	Correct	Falsely Negative

Figure 7. Heat Table of Deviating Results for Caligari.

Through the use of heatmap visualizations (Figures 7-9), we can immediately identify cases where reviews are classified as negative (red), mixed (yellow), or positive (green). Numerical scores follow that scale in a gradient, with normalized values from -1 to +1.

CALIGARI

There were 10 instances (26%) where VADER deviated from our manual analyses of *Caligari*. As we shall see, this was the best result out of the three films. The ChatGPT-VADER model only made two mistakes, and in both cases VADER-only had made the same mistake. In the case of *Caligari* reviews, the direction of mistakes is fairly even – there were four false positives and 6 false negatives in the VADER-only model, and one of each for the two mistakes of the hybrid model (Figure 7). We will proceed to examine these two exceptional cases more closely.

The first case pertains to the review “Ein expressionistischer Film” by Herbert Ihering, published on February 29, 1920, in the *Berliner Börsen-Courier* (ID “PJSEZGQ4”). Despite the manual encodings indicating a negative stance, both the VADER-only and VADER-ChatGPT models assigned the review a positive sentiment. This was one of the few cases in which we believe ChatGPT indubitably made a mistake, rather than displaying an acceptable divergence of judgment from ours, since the reviewer leaves little question as to his negative opinion of the film. Closely examining the review and ChatGPT’s output, it becomes evident that statements originally framed in a negative context, critiquing the untapped potentials of the medium and cinematic expressionism, were interpreted by the model in a positive light :

ChatGPT:

The review can be considered positive. The reviewer acknowledges the significance of expressionism in film and praises the film’s depiction of a “sane reality” being opposed by the idea of “sick unreality”. [...]

This is actually a reversal of the reviewer’s opinion, who criticizes the film’s use of expressionism only because of its setting:

It is telling that Carl Mayer and Hans Janowitz rendered their photoplay *The Cabinet of Dr. Caligari* expressionistically only because it is set in an insane asylum. It opposes the notion of a sick unreality to the notion of a healthy reality. [...] In other words, insanity becomes the excuse for an artistic idea.

Ihering’s argumentation is intricate and he doesn’t belabor his points by stating them explicitly, which would help explain why the model struggled, but it is clear in context that the author is a proponent of expressionism and finds the film’s casual use of it as an aesthetic to signify madness offensive.

The second case of misattribution was of Roland Schacht’s critique published on March 14, 1920, in the *Freie Deutsche Bühne* (ID “VNXHI9SG”), which we manually coded as positive, but was assessed as negative by both VADER alone and VADER-ChatGPT. Interestingly, an inspection of ChatGPT’s response is closer to the mark, suggesting a mixed sentiment:

ChatGPT:

Based on the information provided, it is difficult to determine whether the film review is negative or positive. The review primarily focuses on the plot and artistic elements of the film, such as the set design and performances. However, the reviewer does mention some disappointments with certain aspects of the film’s narrative and stylization. Overall, the review seems

to acknowledge the artistic significance of the film while also highlighting its flaws. Therefore, it can be seen as a balanced review rather than entirely negative or positive.

A potential explanation for the undetected positive essence of the review could lie in the fact that the author's fundamental admiration for *Caligari* is somewhat obscured within the middle portion, following an extensive plot summary. Given the film's subject matter, the plot summary is littered with words that VADER scores as negative, such as murderer (-3.6), horrible (-2.5) and death (-2.9) (vader_lexicon.txt).

It is also true that the reviewer makes some criticisms of the film, such as that its twist ending undermines the audience's initial response, but he follows that criticism with this rhapsody:

But this complaint is as good as negligible in view of the other great artistic significance of this film as a whole. For the first time, the film is fundamentally lifted out of the realm of photography into the pure sphere of the work of art; for the first time, the emphasis is fundamentally placed not on the what of the brutal and exciting events, but on the how; for the first time, not a vulgar illusionistic, but an artistic effect is striven for. (697, ID "VNXHI9SG")

Schacht's recurring rhetorical use of "for the first time" not only highlights *Caligari*'s novelty, but hails it as a watershed moment for film as an artform. This level of effusiveness is a common feature in *Caligari*'s reviews, and is sadly lost in a process of sentiment analysis that weighs things simply as "positive", "mixed" or "negative".

METROPOLIS

Metropolis proved to be the most challenging case for the model, with deviations from the manual analyses in 12 out of 25 reviews, 48% (Figure 8). The model seems to have struggled in particular with *Metropolis* because of the large number of mixed reviews (52%, see Figure 5) that skew negative in the binary assessment. Out of the mistakes the VADER-only model made concerning *Metropolis*, they were *all* misidentifications of the review as positive.

It is little wonder that *Metropolis* was ultimately the film with the largest number of errors, since the movie garnered reviews that seem almost tailor-made to challenge lexicon-based sentiment analysis models. The majority of *Metropolis* reviews commend various technical aspects of the film, but ultimately condemn it for its poor plot and characters. A paradigmatic example is offered by this excerpt from Donald Beaton's review in *The Film Spectator* (ID "KIUF2ZGM"):

ID	Movie	Manual_Judgment	Binary_Judgment	VADER_ChatGPT_sentiment	VADER_ChatGPT_score	VADER_review_sentiment	VADER_review_score	GPT_Error_Check	VADER_Error_Check
KIUF2ZGM	Metropolis	mixed	negative	positive	0.502300	positive	0.981100	Falsely Positive	Falsely Positive
EIRLBFNR	Metropolis	mixed	negative	positive	0.519200	positive	0.993800	Falsely Positive	Falsely Positive
JUZY74XR	Metropolis	mixed	negative	negative	-0.745400	positive	0.987700	Correct	Falsely Positive
KCDMVU3Q	Metropolis	negative	negative	negative	-0.955200	positive	0.996000	Correct	Falsely Positive
TIRHYW52	Metropolis	mixed	negative	mixed	-0.063400	positive	0.927400	Correct	Falsely Positive
AXSMZ7XF	Metropolis	mixed	negative	positive	0.518700	positive	0.992400	Falsely Positive	Falsely Positive
C5HFH5UD	Metropolis	negative	negative	negative	-0.829600	positive	0.902700	Correct	Falsely Positive
5KUVWUIP	Metropolis	negative	negative	negative	-0.920100	positive	0.995100	Correct	Falsely Positive
ZV2KYKDS	Metropolis	mixed	negative	negative	-0.901300	positive	0.997700	Correct	Falsely Positive
D4AR9KYS	Metropolis	negative	negative	positive	0.981000	positive	0.997800	Falsely Positive	Falsely Positive
CZWZ5UV7	Metropolis	mixed	positive	negative	-0.700300	positive	0.996100	Falsely Negative	Correct
TM6H4KLS	Metropolis	mixed	negative	negative	-0.868100	positive	0.998000	Correct	Falsely Positive

Figure 8. Heat table of Deviating Results for Metropolis.

Technically, *Metropolis* is a great picture. The sets were marvelous, and the mob direction was good. The whole picture showed wonderfully painstaking care in the production, but good production is not enough to make a good picture.

This sentence presents four positive aspects, yet the significance of the negative aspect clearly outweighs them. Lexicon-based sentiment analysis is fundamentally incapable of capturing such nuance, since it offers the sentiment of a text by adding up – with some small adjustments – the individual sentiment scores of words and then averaging them out. In a case like Beaton's review, the high quantity of positive words will ultimately deliver a positive result, despite the author's clear negative feelings about the picture as a whole.

In [Figure 8](#), we can see that both the VADER-only method and the hybrid approach wrongly categorized Beaton's rather scathing review as positive. If we inspect ChatGPT's output, however, we find it to be a remarkably accurate summary of Beaton's – mostly negative – thoughts:

The review is negative. The reviewer acknowledges the technical aspects of the film, such as the marvelous sets and good direction, but criticizes the lack of story and mediocre acting. The reviewer also states that the idea of the film was wrong and points out inconsistencies in the plot. They ultimately attribute the credit for the film's success to the technicians and the person responsible for the set design and machinery.

It is worth reiterating that both of the methods we employed use VADER, so the flaws of a lexicon-based sentiment analysis are present in both of them. Our implementation of ChatGPT as an intermediate step served simply to alleviate this problem of naive quantification, but it did not solve it. It is worth noting, however, that the hybrid model did manage to lower the score of Beaton's review from an astonishingly high 0.98 to a moderately

ID	Movie	Manual_Judgment	Binary_Judgment	VADER_ChatGPT_sentiment	VADER_ChatGPT_score	VADER_review_sentiment	VADER_review_score	GPT_Error_Check	VADER_Error_Check
8WNFN9AE	Nosferatu	negative	negative	negative	-0.971300	positive	0.869600	Correct	Falsely Positive
5YA546KU	Nosferatu	mixed	positive	negative	-0.880700	positive	0.666600	Falsely Negative	Correct
9W4Z3TFU	Nosferatu	mixed	positive	positive	0.928700	negative	-0.995600	Correct	Falsely Negative
K7EEHSC6	Nosferatu	mixed	positive	positive	0.931300	negative	-0.891600	Correct	Falsely Negative
CEHEIWR7	Nosferatu	mixed	positive	negative	-0.911800	positive	0.578900	Falsely Negative	Correct
2SQ2YDPA	Nosferatu	mixed	positive	positive	0.739400	negative	-0.980200	Correct	Falsely Negative
KWC9KUTP	Nosferatu	negative	negative	negative	-0.755500	positive	0.994100	Correct	Falsely Positive
L7JBS7LN	Nosferatu	positive	positive	positive	0.943300	negative	-0.969200	Correct	Falsely Negative

Figure 9. Heat Table of Deviating Results for Nosferatu.

positive 0.50, and similar reductions happened in other cases (Beaton, ID “KIUF2ZGM”; Anonymous, “Production”, ID “EIRLBFNR”; Haas, ID “AXSMZ7XF”).

A closer inspection of these instances show that ChatGPT’s responses were also mischaracterized by VADER, as in all three cases ChatGPT correctly and unambiguously categorized the reviews as negative. The chatbot’s analysis of the other two reviews where the hybrid model diverged from the manual encoding (Bartlett *apud* Minden and Bachmann ID “D4AR9KYS”; Barry, ID “CZWZ5UV7”) still display a level of nuance that makes calling them “incorrect” feel disingenuous. While their judgments differ from our manual judgments, we do not substantially disagree with the LLM’s outputs, finding that they could well have been written by another human who simply weighed the reviews slightly differently than we did.

NOSFERATU

With 8 out of 17 (47%), almost half of film reviews for *Nosferatu* were inaccurately assessed by one of the two models (Figure 9). As was the case with *Metropolis*, this mostly pertains to reviews classified as mixed in our original manual analysis.

Interestingly, the misassessments of the automated methods do not align in any of the cases. While VADER incorrectly labeled four reviews as negative and two reviews as positive, VADER-ChatGPT identified two reviews as false negatives. However, in both of those cases, the initial manual analysis was originally mixed and was only categorized as positive through the enforced binary classification.

In contrast to the errors in *Metropolis*, most misidentifications (6 out of 8) in the reviews of *Nosferatu* erred on the negative side. Admittedly, even those reviews often carried significantly negative comments (which is why so many of them had initially been labeled as mixed). Upon closer examination of the two outcomes marked as incorrect negatives in the VADER-ChatGPT sentiment analysis, the intricacies of each case become evident. This complexity is also intertwined with the intended readership of each review.

In the review “Film Reviews - Nosferatu the Vampire,” published on December 12, 1929, in *Variety* (ID “5YA546KU”), the anonymous author conveys a combination of sentiments regarding the anticipated success of *Nosferatu*. On the one hand, the review highlights the film’s directorial and artistic aspects, primarily crediting Murnau’s directing talent and consistently attributing the “[s]killfully mounted and directed” (26) work to him:

Murnau proved his directorial artistry in “Sunrise” for Fox about three years ago,¹⁸ but in this picture he’s a master artisan demonstrating not only a knowledge of the subtler side of directing but in photography. One shot of the sun cracking at dawn is an eye filler. Among others of extremely imaginative beauty is one which takes in a schooner sailing in a rippling stream photographed in such a manner that it has the Illusion of color and an enigmatic weirdness that’s more perplexing than the ghost action of the players. (Ibid.)

On the other hand, the review also stresses some shortcomings, especially concerning titling, acting, and the adaptation of Bram Stoker’s *Dracula* for the screenplay. It characterizes *Nosferatu* as “a depressive piece of art,” and raises concerns about its potential unsuitability for mainstream theatrical audiences. One aspect that must be considered when analyzing (historical) film reviews is that not all of them have the same objectives. This review was published in *Variety*, a trade journal, and its language, particularly its emphasis on being a “risky exhibit,” (26) appears to be directed specifically at cinema owners.

In our manual judgment, we found this review to be mixed, but leaning positive when pressed because the reviewer does emphasize the film’s artistic merits and overall quality. It is interesting to note that this review was so ambivalent that ChatGPT’s itself went back and forth on its judgment. Its assessment of the review when we used it as an assistant for the binarization process¹⁹ was that the review was mostly positive – a judgment we agreed with at the time. Ultimately, this case is so ambivalent that both of its responses are in the realm of acceptable subjective divergence which another human coder might also make.

¹⁸ Though *Nosferatu* was released in Germany in 1922, it was only released in the US in 1929, buoyed in part by Murnau’s success with his American-made *Sunrise* in 1927. By the time of *Nosferatu*’s release in the US, sound film had already begun taking over the screens, which certainly did not help the film’s American reviews, which mostly - as is the case here - seemed unaware that it was an older film and found it to be a step back for the director.

¹⁹ See column “ChatGPT Binary Answers”.

In the second instance where VADER-ChatGPT incorrectly attributed a negative sentiment, the distinction between positive and negative attitudes was also unclear. Robert Desnos' published this snippet of a review of *Nosferatu* in May 1929 in the journal *La Revue du cinéma* (ID "CEHEIWR7"):

However, an admirable film, far superior to *Caligari* in terms of direction, went unnoticed: *Nosferatu the Vampire*, where no innovation was arbitrary, and where everything was sacrificed to poetry and nothing to art. But, driven by the will of directors oblivious to their role, German artists, obliged to fill an absurd landscape with derisory gestures, soon gave them ridiculous importance.

Overall, the reviewer has a positive view of *Nosferatu* in terms of direction and its focus on poetry but criticizes the perceived lack of innovation and the actions of German artists in the film. What carries more weight in this context, the positive or the negative statements? And what serves as an indicator? In the manual analysis, the tendency towards positivity is attributed to the first sentence, incorporating markers like "[h]owever" and "went unnoticed," which aims to communicate to readers that the film was undeservedly overlooked by a wider audience – similar to what's stated in the *Variety* review. The follow-up sentence, however, is undeniably negative. At first glance, this seems to be a critique of *Nosferatu*, but careful consideration indicates that the criticism is targeted at *other* German films in the *Caligari* mold, which the author dislikes for (unlike *Nosferatu*) allegedly displaying arbitrary innovation and absurd landscapes. This level of intricate argumentation is even difficult for a human to parse, so it is little wonder that ChatGPT struggled with it.

Conclusion

As we demonstrated above, determining the sentiment of intricate and nuanced film reviews can be a challenge, even for human experts with domain knowledge. This underscores the limitations of adopting a simplistic (binary) approach to sentiment. Film scholars used to qualitative research may well question the utility of such an endeavor. We find that the value is two-fold. The first is the testing of hypotheses that have already been developed with close reading, so as to provide stronger evidence in favor of them, or as to refute them. For instance, most people with even a passing knowledge of *Metropolis*' contemporary reception could tell you that critics found the film spectacular in its technical aspects, but ridiculous in its plot and social message. Establishing that 52% of *Metropolis* reviews were mixed is but a confirmation of that – mundane, perhaps, but still worthwhile. This application certainly feels more useful when there is scholarly debate about a film's reception, such as in the case of *Caligari*, sketched in the introduction. Our analysis showed that though not unanimous, the film was mostly (76%)

received with glowing reviews, and the positive reception could be seen across countries and publication types. The second – and more important – value of distant reading methods is when it comes to large corpora. This would allow us to get away from the canon and make judgments about films that few – if any – scholars would otherwise dedicate themselves to researching through the use of close reading. Given a clean corpus, we could answer, with a reasonable degree of certainty, large questions such as “were German movies well received internationally?” and we could break down their reception in several countries without focusing only on a small and skewed sample of films.

It is important to acknowledge that access to large datasets of clean historical sources that can be seamlessly integrated in such automated processes can be the tallest hurdle when it comes to implementing such methods. Though digital methods in general and LLMs in particular are routinely touted for their potential time-saving advantages, this is often a false promise. While most other papers that undertake sentiment analysis of film reviews utilize born-digital corpora like IMDb reviews, we went through considerable lengths to (manually) obtain, clean and categorize digitized historical sources that were diverse in language, country of origin and publication type. Identifying the correct films in full-text search, distinguishing between mere mentions of films, advertisements, and actual reviews, defining the boundaries of individual reviews on the page and handling the quality issues of OCR were all extremely time-consuming, even for our ultimately small corpus of 60 reviews.

We believe that current technological developments in the realm of machine learning will continue to yield gains in those areas, increasing, for instance, the quality of OCR and facilitating the identification of texts related to particular movies. These developments are not only helpful for quantitative analysis (sometimes wrongly seen as synonymous with “digital methods”), but also for qualitative analysis. In addition to better quality of textual data and search capabilities, close reading itself is a method that can gain from interaction with LLMs. We were impressed with ChatGPT’s ability to reliably summarize the main arguments of a review, and felt that even in the instances in which we disagreed with its responses, it was not out of the realm of possibility for a human to have reached those same conclusions. Although LLMs are unlikely to offer entirely novel or exceptionally insightful analyses, they can serve as valuable sparring partners and as easily implementable inter-rater reliability measures. While at the point of writing open source alternatives like HuggingChat did not yet match the capabilities of private models, we are optimistic about the potential for such models to improve over time.

Despite the impressive capabilities of LLMs – proprietary or not – to handle nuanced text, these models proved themselves to be unreliable for tasks requiring consistency and reproducibility. When it comes to sentiment analysis, lexicon-based models deliver such consistency, but are unfortunately highly inaccurate. The contrasting advantages and disadvantages of these two methods led us to propose a combination of the two, running a lexicon-based model not on the film reviews directly, but on the analysis of those reviews made by ChatGPT. It is worth reiterating that our method is still ultimately a lexicon-based model, so the weaknesses of that approach are still very much present. The model struggles in particular with a corpus such as ours, which contains film reviews with an abundance of negatively coded words (“murder”, “horrifying,” etc.) without them necessarily indicating a negative review of the film. This is likely one of the reasons why *Metropolis* – a visually dazzling science fiction film – had so many reviews that were wrongly identified as positive, in contrast to the other two movies, which are pioneering examples of the horror genre.

Despite the challenges provided by such a corpus, our hybrid model, combining ChatGPT’s nuanced handling of language with lexicon-based sentiment analysis, delivered respectable results. We believe that combining machine learning methods with symbolic (human-readable) programming is a promising avenue when it comes to automating complex tasks, and we expect further improvements in this area in the coming years. While there are inherent challenges associated with adopting a digital approach to the analysis of historical film-related texts, we firmly believe that the benefits it offers are substantial. While not mandatory, the use of digital methods for analyzing film reviews holds the potential to significantly enhance the quality of evidence, especially when dealing with larger datasets. It’s important to note that these digital methodologies are not meant to replace qualitative approaches, but to complement them. By broadening our scope to include a more diverse range of film reviews, extending beyond the foundational elements of the Weimar film canon, we can more effectively address specific comparative questions.

Furthermore, the process of constructing and sharing an explicit corpus has compelled us to be exceptionally precise and transparent about our methodology – a level of detail not commonly encountered in the writing of qualitative historical papers. Although this paper primarily focuses on generating quantitative insights, such as the percentage of positive reviews, the sharing of our complete dataset and code enables others to critique and leverage our work for various purposes, including qualitative analyses. We hope the methodology outlined in this article will serve as a framework for those interested in automated analysis of film reviews and inspire further research at the intersection of digital methods and film history.

Data repository: <https://doi.org/10.7910/DVN/8NINQK>

Submitted: March 20, 2024 EDT, Accepted: May 21, 2024 EDT



This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-4.0). View this license's legal deed at <http://creativecommons.org/licenses/by/4.0> and legal code at <http://creativecommons.org/licenses/by/4.0/legalcode> for more information.

WORKS CITED

- Acland, Charles R., and Eric Hoyt. *The Arclight Guidebook to Media History and the Digital Humanities*. Reframe, 2016.
- Amulya, K., et al. "Sentiment Analysis on IMDB Movie Reviews Using Machine Learning and Deep Learning Algorithms." *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, 2022, pp. 814–19. DOI.org (Crossref), <https://doi.org/10.1109/ICSSIT53264.2022.9716550>.
- Andriopoulos, Stefan. "Suggestion, Hypnosis, and Crime: Robert Wiene's *The Cabinet of Dr. Caligari* (1920)." *Weimar Cinema: An Essential Guide to Classic Films of the Era*, edited by Noah Isenberg, Columbia University Press, 2009, pp. 13–32.
- Anonymous. "Film Reviews - Nosferatu the Vampire." *Variety*, vol. XCVII, no. 11, Dec. 1929, pp. 26–30.
- . "Production Weak in Human Qualities." *The Film Spectator*, vol. 4, no. 1, Sept. 1927, p. 5.
- Arnold, Taylor, and Lauren Tilton. "Distant Viewing: Analyzing Large Visual Corpora." *Digital Scholarship in the Humanities*, vol. 34, no. Supplement_1, Dec. 2019, pp. i3–16. DOI.org (Crossref), <https://doi.org/10.1093/llc/fqz013>.
- Ascari, Maurizio. "The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres." *Genre*, vol. 47, no. 1, Apr. 2014, pp. 1–19. DOI.org (Crossref), <https://doi.org/10.1215/00166928-2392348>.
- Barlow, John D. *German Expressionist Film*. Twayne, 1982.
- Barry, Iris. "The Cinema: Metropolis." *The Spectator*, no. 540, 26 Mar. 1927.
- Beaton, Donald. "As They Appeal to a Youth." *The Film Spectator*, vol. 4, no. 4, Oct. 1927, p. 17.
- Bukar, Umar, et al. "Text Analysis of Chatgpt as a Tool for Academic Progress or Exploitation." *Preprint*, SSRN, 2023. DOI.org (Crossref), <https://doi.org/10.2139/ssrn.4381394>.
- Chen, Cheng, et al. "Sentiment Analysis of Animated Film Reviews Using Intelligent Machine Learning." *Computational Intelligence and Neuroscience*, edited by Qiangyi Li, vol. 2022, July 2022, pp. 1–8. DOI.org (Crossref), <https://doi.org/10.1155/2022/8517205>.
- Ciotti, Fabio. "Distant Reading in Literary Studies: A Methodology in Quest of Theory." *Testo e Senso*, no. 23, Dec. 2021, pp. 195–213.
- Denbo, Seth, and Neil Fraistat. "Diggable Data, Scalable Reading and New Humanities Scholarship." *2011 Second International Conference on Culture and Computing*, 2011, pp. 169–70. *IEEE Xplore*, <https://doi.org/10.1109/Culture-Computing.2011.49>.
- Desnos, Robert. "Scénarios." *La Revue du cinéma*, no. 3, May 1929.
- Elsaesser, Thomas. *Weimar Cinema and After: Germany's Historical Imaginary*. Digital Edition, Routledge, 2009.
- Farias, D. I. Hernández, and P. Rosso. "Irony, Sarcasm, and Sentiment Analysis." *Sentiment Analysis in Social Networks*, Elsevier, 2017, pp. 113–28. DOI.org (Crossref), <https://doi.org/10.1016/B978-0-12-804412-4.00007-3>.
- Flueckiger, Barbara, and Gaudenz Halter. "Methods and Advanced Tools for the Analysis of Film Colors in Digital Humanities." *Digital Humanities Quarterly*, vol. 14, no. 4, 2020, <https://doi.org/10.5167/uzh-197994>.
- Hariri, Walid. *Unlocking the Potential of ChatGPT: A Comprehensive Exploration of Its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing*. 2023. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2304.02017>.

- Howanitz, Gernot. "Distant Watching. Ein quantitativer Zugang zu YouTube-Videos." *Proceedings of the 2nd Annual Conference of the Association 'Digital Humanities Im Deutschsprachigen Raum' (DHD 2015), Graz, 23.-27 Feb. 2015.*
- Hoyt, Eric. *Ink Stained Hollywood: The Triumph of American Cinema's Trade Press.* University of California Press, 2022.
- Hutto, C., and Eric Gilbert. "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 8, no. 1, May 2014, pp. 216–25. *ojs.aaai.org*, <https://doi.org/10.1609/icwsm.v8i1.14550>.
- Hutto, C. J. *vaderSentiment/Vader_lexicon.Txt*. 1 Apr. 2022, https://github.com/cjhutto/vaderSentiment/blob/master/vaderSentiment/vader_lexicon.txt.
- Ihering, Herbert. "Der Metropolisfilm - Ufa-Palast am Zoo." *Berliner Börsen-Courier*, Late edition, 11 Jan. 1927, p. 2.
- . "Der Metropolisfilm - Ufa-Palast am Zoo (Vorbericht)." *Berliner Börsen-Courier*, Early edition, 11 Jan. 1927, p. 5.
- . "Ein expressionistischer Film." *Berliner Börsen-Courier*, Early edition, no. 101, 29 Feb. 1920, p. 8.
- Kaes, Anton, et al., editor. *The Weimar Republic Sourcebook*. 1., Univ. of California Press, 1995, <https://doi.org/10.1525/9780520909601>.
- Kracauer, Siegfried. *From Caligari to Hitler: A Psychological History of the German Film*. Edited by Leonardo Quaresima, Princeton University Press, 2004.
- Landry, Lionel. "Le Cabinet du Docteur Caligari." *Cinéa*, no. 27, Nov. 1921, p. 8.
- Lu, Kui, and Jiesheng Wu. "Sentiment Analysis of Film Review Texts Based on Sentiment Dictionary and SVM." *Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence*, ACM, 2019, pp. 73–77. DOI.org (Crossref), <https://doi.org/10.1145/3319921.3319966>.
- Maas, Andrew L., et al. "Learning Word Vectors for Sentiment Analysis." *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, <https://ai.stanford.edu/~ang/papers/acl11-WordVectorsSentimentAnalysis.pdf>.
- Maynard, Diana, and Mark Greenwood. "Who Cares about Sarcastic Tweets? Investigating the Impact of Sarcasm on Sentiment Analysis." *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), 2014, pp. 4238–43, http://www.lrec-conf.org/proceedings/lrec2014/pdf/67_Paper.pdf.
- Minden, Michael, and Holger Bachmann, editors. *Fritz Lang's Metropolis: Cinematic Visions of Technology and Fear*. Camden House, 2002.
- Moretti, Franco. *Distant Reading*. Verso, 2013.
- Open Syllabus. "OS Movie Lab." *Open Syllabus*, 2021, <https://blog.opensyllabus.org/movie-lab/>.
- Orndorff, Marguerite. "[25] METROPOLIS (Paramount)." *The Educational Screen*, vol. VII, no. 2, Apr. 1928, p. 68.
- Primorac, A., et al. "Distant Reading Two Decades on: Reflections on the Digital Turn in the Study of Literature." *Digital Studies/Le Champ Numerique*, 2023. *ora.ox.ac.uk*, <https://ora.ox.ac.uk/objects/uuid:acbba668-71d1-4f8e-9cfc-fda801214682>.
- Rebora, Simone. "Sentiment Analysis in Literary Studies. A Critical Survey." *Digital Humanities Quarterly*, vol. 017, no. 2, July 2023.
- Robinson, David. *Das Cabinet des Dr. Caligari*. BFI Pub, 1997.
- Schacht, Roland. "Caligari." *Freie Deutsche Bühne (Berlin)*, no. 29, 14 Mar. 1920, pp. 695–98.

- Shaukat, Zeeshan, et al. "Sentiment Analysis on IMDB Using Lexicon and Neural Networks." *SN Applied Sciences*, vol. 2, no. 2, Feb. 2020, p. 148. DOI.org (Crossref), <https://doi.org/10.1007/s42452-019-1926-x>.
- Shen, Yongliang, et al. *HuggingGPT: Solving AI Tasks with ChatGPT and Its Friends in Hugging Face*. 2023. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2303.17580>.
- Sight and Sound. "The 100 Greatest Films of All Time." *BFI*, 28 June 2021, <https://www.bfi.org.uk/sight-and-sound/polls/greatest-films-all-time-2012>.
- Sudirjo, Frans, et al. "Application of ChatGPT in Improving Customer Sentiment Analysis for Businesses." *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 5, no. 3, July 2023, pp. 283–88. DOI.org (Crossref), <https://doi.org/10.47233/jteksis.v5i3.871>.
- Thompson, Kristin. "Dr. Caligari at the Folies-Bergère, or, The Successes of an Early Avant-Garde Film." *The Cabinet of Dr. Caligari: Texts, Contexts, Histories*, edited by Mike Budd, Rutgers University Press, 1990.
- Topal, Kamil, and Gultekin Ozsoyoglu. "Movie Review Analysis: Emotion Analysis of IMDb Movie Reviews." *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016, pp. 1170–76. *IEEE Xplore*, <https://doi.org/10.1109/ASONAM.2016.7752387>.
- Tsvian, Yuri, and Gunars Civjans. *Cinematics: Movie Measurement and Study Tool Database*. 2005, <https://cinematics.uchicago.edu/database.php>.
- Underwood, Ted. "A Genealogy of Distant Reading." *Digital Humanities Quarterly*, vol. 011, no. 2, June 2017.
- Wankhade, Mayur, et al. "A Survey on Sentiment Analysis Methods, Applications, and Challenges." *Artificial Intelligence Review*, vol. 55, no. 7, Oct. 2022, pp. 5731–80. *Springer Link*, <https://doi.org/10.1007/s10462-022-10144-1>.
- Ziems, Caleb, et al. *Can Large Language Models Transform Computational Social Science?* 2023. DOI.org (Datacite), <https://doi.org/10.48550/ARXIV.2305.03514>.