Informatics

CommentSense

Eine lokal laufende KI Browser Erweiterung zur Echtzeitanalyse von YouTube Kommentaren

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Marc Kroll, BSc
Matrikelnummer 01302580

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Assistant Prof. Dr.in phil. Mag.a phil. Astrid Weiss
Mitwirkung: Dipl.-Ing. Rafael Vrecar, BSc

Wien, 31. März 2025

[redacted] Marc Kroll, BSc

[redacted] Astrid Weiss

TU WIEN Informatics

CommentSense

An On-Device AI Browser Extension for Real-Time YouTube Comment Understanding

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Marc Kroll, BSc
Registration Number 01302580

to the Faculty of Informatics

at the TU Wien

Advisor: Assistant Prof. Dr.in phil. Mag.a phil. Astrid Weiss
Assistance: Dipl.-Ing. Rafael Vrecar, BSc

Vienna, March 31, 2025

[redacted]
Marc Kroll, BSc

[redacted]
Astrid Weiss

Erklärung zur Verfassung der Arbeit

Marc Kroll, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.
Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 31. März 2025

[redacted]

Marc Kroll, BSc

Danksagung

Ich möchte meinen Dank dem österreichischen Steuerzahler aussprechen, der es mir durch die Möglichkeit der Bildungskarenz ermöglicht hat, diese Arbeit und damit auch das Studium erfolgreich abzuschließen.

Ein besonderer Dank gilt meiner Partnerin, Verena Ingrid Prantl, für ihre unermüdliche Unterstützung und Motivation, die mir während des gesamten Prozesses Kraft und Zuversicht gegeben hat.

Ebenso danke ich meinen Freunden und Kollegen, die mich über die Jahre hinweg davon abgehalten haben, das Studium abzubrechen, und mich stattdessen immer wieder ermutigt haben, meine Masterarbeit endlich zu schreiben. Jetzt ist es soweit: Ich habe es geschafft—Dank' euch allen!

Mein Dank richtet sich auch an meine betreuende Professorin und ihren Assistenten für das entgegengebrachte Vertrauen, die fachliche Unterstützung und das wertvolle, vor allem positive und motivierende Feedback, das maßgeblich zum Gelingen dieser Arbeit beigetragen hat.

Acknowledgements

I would like to express my gratitude to the Austrian taxpayers, who made it possible for me to successfully complete this thesis—and thus my studies—through the opportunity of *educational leave*.

A special thanks goes to my partner, Verena Ingrid Prantl, for her tireless support and motivation, which gave me strength and confidence throughout the entire process.

I am also deeply grateful to my friends and colleagues who, over the years, kept me from dropping out and instead continuously encouraged me to finally write my masters thesis. Now the moment has come: I did it—thanks to all of you!

My sincere thanks also go to my supervising professor and her assistant for their trust, professional guidance, and valuable feedback, which was especially positive and motivating and played a key role in the success of this work.

Kurzfassung

YouTube, das zweitgrößte soziale Netzwerk mit über 2,5 Milliarden monatlichen Nutzern, steht vor erheblichen Herausforderungen im Bereich der Content-Moderation. In der ersten Jahreshälfte 2024 wurden mehr als 2,8 Milliarden Kommentare entfernt, wobei jedoch nur 2,1% hasserfüllte oder beleidigende Inhalte waren. Diese Diskrepanz verdeutlicht die Grenzen automatisierter Systeme und führt dazu, dass menschliche Moderatoren täglich mit den dunkelsten Seiten des Internets konfrontiert werden.

Die Sentiment-Analyse von YouTube-Kommentaren bietet eine vielversprechende Lösung, indem sie es Nutzern ermöglicht, bewusst mit positivem, neutralem oder negativem Inhalt zu interagieren. Obwohl BERT-basierte Modelle wie DistilBERT-sst2 gute Ergebnisse liefern, ist unklar, ob sie von kleinen Sprachmodellen (Small Language Models, SLMs) übertroffen werden können—insbesondere, da SLMs Kontext besser einbeziehen und darüber hinaus erklären können, welche Inhalte für das Klassifikationsergebnis relevant waren.

Diese Studie verfolgt einen Mixed-Methods-Ansatz, der mit einer Nutzerpräferenzstudie (prototypbasierte Interviews) beginnt, um erste Einblicke zu gewinnen. Darauf folgten ein Pilotversuch und eine umfassende Nutzerevaluation, einschließlich Umfragen, einem Aufgabenblatt und Interviews. Abschließend wurde eine Modellauswertung durchgeführt, um die Übereinstimmung der Modelle mit einer zuvor festgelegten menschlichen *Ground-Truth-Basis* zu beurteilen. Der Pilotversuch vor der Nutzerevaluation zeigte, dass Ansätze mit vordefinierten Filterkriterien, wie Gruppierung, Sortierung oder Hervorhebung von Kommentaren, schlecht aufgenommen wurden, da sie als wenig nützlich empfunden wurden. Stattdessen äußerten die Teilnehmer den Wunsch, eigene Filter setzen zu können und nur bei Bedarf zusätzliche Informationen über das automatisierte Labeling-System abzurufen. Basierend auf diesen Erkenntnissen wurden drei Browser-Erweiterungen zur Sentiment-Klassifikation entwickelt: Tool A (nutzt DistilBERT-sst2), Tool B (verwendet ein SLM mit vereinfachten Klassifikationen) und Tool C (nutzt ein SLM mit einem mehrstufigen Klassifikationssystem). In der Nutzerstudie wurden diese Tools miteinander verglichen, wobei das SLM-basierte Vorgehen in der Nutzerpräferenz konsistent besser abschnitt als DistilBERT-sst2. Dies lag vor allem daran, dass die SLMs die Klassifikationsentscheidungen begründeten. Diese Transparenz führte zu einem höheren Vertrauensgefühl gegenüber den SLM-basierten Tools bei den Nutzern. Darüber hinaus übertrafen 80% der bewerteten SLMs DistilBERT-sst2 in einer Zero-Context-

Binärklassifikation auf dem YouTube-Kommentar-Datensatz, und 65% waren auch in kontextbasierten Binärklassifikationen überlegen. Insgesamt erzielten die leistungsstärksten Modelle höhere Macro-F1-Scores, wenn sie durch kontextuelle Informationen ergänzt wurden, und näherten sich so stärker der menschlichen *Ground-Truth* an.

Durch den Einsatz lokal laufender SLMs bietet diese Forschung einen neuartigen Ansatz zur Reduzierung schädlicher Inhalte in sozialen Medien, während gleichzeitig die Nutzerautonomie erhalten bleibt—das heißt, die Nutzer können selbst entscheiden, welche Inhalte sie sehen möchten. Als Browser-Erweiterung birgt ein solches Tool großes Potenzial, Nutzer dazu zu befähigen gezieltere Online-Interaktionen zu haben.

Abstract

YouTube, the second-largest social network with over 2.5 billion monthly users, faces significant challenges in content moderation. In the first half of 2024, over 2.8 billion comments were removed, but only 2.1% targeted hateful or abusive content. This gap highlights the limitations of automated systems, placing a heavy burden on human moderators who face the internets darkest content daily. Sentiment analysis of YouTube comments offers a promising solution, enabling users to consciously engage with positive, neutral, or negative content. While BERT-based models like DistilBERT-sst2 perform well, it remains unclear whether they can be outperformed by small language models (SLMs), especially given that SLMs are able to incorporate and reason about context, a potential benefit, when classifying YouTube comments. This study used a mixed-methods approach, beginning with a user preference study (prototype-based interviews) to gather initial insights. This was followed by pilot testing and a comprehensive user evaluation, including surveys, user tasks, and interviews. Finally, a model evaluation assessed the alignment of the models against a prior established ground truth. The user preference study prior to the user evaluation study showed that approaches with predefined filtering criteria, such as comment grouping, reordering, or highlighting, were poorly received because they were not considered useful. Instead, participants expressed a preference for setting their own filters to interact with an automated labeling system that only provides additional information when demanded. Based on these insights, three browser extensions for sentiment classification were developed: Tool A (uses DistilBERT-sst2), Tool B (uses an SLM with simplified classifications), and Tool C (uses an SLM with a multi-tier classification system). In the user evaluation study, these tools were compared, revealing that the SLM-based approach outperformed DistilBERT-sst2 consistently in user preference. This was primarily due to the SLMs ability to provide reasoning for classification decisions, offering higher perceived trustworthiness due to transparency. Additionally, 80% of the evaluated SLMs outperformed DistilBERT-sst2 in zero-context binary classification on the YouTube comment dataset, and 65% did so in context-based binary classification. Overall, the top-performing models achieved higher macro-F1 scores when augmented with contextual information, demonstrating closer alignment with human judgment. By employing locally running SLMs, this research offers a novel approach to reducing harmful content in social media while maintaining user agency—allowing users to decide what they see. As a browser extension, such a tool has high potential for empowering users and fostering healthier online interactions.

Contents

CHAPTER 1

Introduction

As of April 2024, YouTube is the second-largest social network worldwide with over 2.5 billion active monthly users, only surpassed by facebook (3 billion) [16]. In September 2024, YouTube ranked as the second most visited website globally right behind Google [104]. On YouTube, videos can be uploaded, shared, liked or disliked, and commented. In the first half of 2024 over 2.8 billion YouTube video comments were removed due to violations of the platform's community guidelines—majorly through automated flagging. For comparison, in all of 2023 roughly 3.6 billion comments were removed [18]. Looking a bit closer, one can see that 80.8% of the removed comments are attributed to being spam, misleading or scam content, 6.9% to harassment and cyberbullying, 5% to endangering child safety, 4.3% to violent or graphic content, 2.1% to hateful or abusive comments, and 0.9% to comments promoting violence and violent extremism [17].

These numbers suggest that automated moderation systems are effective at filtering out spam and scams but struggle with detecting toxic, offensive, and harmful comments. Content labeling is one of the strategies platforms use to organize, moderate, and regulate content, alongside removal and algorithmic sorting [78]. Labeling helps platforms provide users with information about the quality of the content they encounter, with labels being either negative or positive. Negative labels indicate low or questionable quality, such as being unverified, false, disputed, or from an unreliable source, while positive labels suggest that the content meets quality standards, such as being verified, fact-checked, or sourced from trusted entities [118]. When deciding to apply labels to content, moderators must carefully consider the labels' intended purpose, effectiveness, and any potential unintended consequences [78].

In their early years, platforms like facebook and YouTube relied on relatively small teams to review content, and their moderation policies were fairly limited in scope [67]. Over time, growing public pressure to remove offensive content led these platforms to develop increasingly sophisticated moderation systems. In 2016, facebook introduced an automatic detection software for its content. For large platforms like YouTube and

Twitter [1], AI could help quickly detecting and removing harmful content, reducing the need for constant human oversight. Ideally, these systems could identify and block content like hate speech or pornography on their own, without human involvement [46, 44]. Still, many social media platforms operate on a feed-based model, using algorithms that curate content based on users' past interactions [46]. Regardless, the specific methods used and strategies behind curating largely remain opaque [57, 67].

Today, most platforms use a hybrid approach to remove problematic content, where algorithms automatically flag some content, and other content is reviewed only after being flagged by users [115]. Human moderators typically review flagged content, though some platforms have recently shifted to almost entirely automated review, particularly in response to the COVID-19 pandemic [76]. While content moderation is intended to be beneficial, it can harm moderators, and the decisions made can negatively affect users [115].

Human content moderators, often employed as crowd-workers [2], remain largely invisible, both by design and circumstance. Many are based in low-wage regions like the Philippines and India, distanced from the platforms they moderate and the users they oversee. Their roles are further obscured by contract labor and intermediary crowd-work platforms. Beyond their hidden status, they face a relentless stream of content—ranging from mistakenly flagged posts to ambiguous cases and graphic human atrocities [46]. Moreover, an active line of research has investigated the "emotional labor" of moderation work by volunteer moderators, further highlighting the importance of avoiding burnout for moderators through automation, as they are exposed daily to the Internets most disturbing content [46, 69].

Meta, for example, has ended partnership with its fact-checking partners in 2025 [66, 15], relying increasingly on automated systems. In his 2018 Congressional testimony, facebook CEO Mark Zuckerberg frequently cited AI as the future solution to the platform's political issues. While it may seem ambitious, the statistics in press releases and company transparency reports show that automation already plays a significant role in enforcing content policies [48]. For example, YouTube reports that 98% of violent extremism videos are flagged by machine learning, and Twitter stated that 93% of accounts removed for terrorist propaganda were flagged by its spam-fighting tools [48, 45]. While AI-driven moderation can handle large volumes of content, it lacks the contextual understanding and nuance that human moderators provide, leading to both over-filtering and the failure to detect harmful speech.

Understanding content moderation is crucial for maintaining a safe online environment and recognizing the impact of social media on public life [46]. The growing use of opaque and powerful algorithms raises questions about how much users should know about their

[1]Twitter is called X since July 2023 [79]. Nevertheless, during this thesis, the term Twitter will be used for consistency with the cited literature.

[2]Crowd-workers are individuals who perform small tasks or jobs, often online, that are part of a larger project.

existence and operation. Whether accurate or not, users' perceptions of these algorithms can influence their behavior [34].

While Online Social Networks (OSNs) rely on content moderation [40], it remains unclear to researchers and platforms themselves how this impacts users [53]. Although algorithms continue to improve in terms of accuracy, fairness, and normative appropriateness, research shows that they are perceived as less legitimate compared to expert panels [88]. Given these concerns about transparency and trustworthiness of these systems, it is important to explore alternative approaches that empower users in the moderation process. This highlights the need of developing a solution that enables users to define their own filtering criteria, thereby giving back control over the content they see, which is also presented in this thesis (see Section: 5.2.2). To the best of my knowledge, no current solution allows users to set their own criteria [3] for content moderation.

In recent years, both the public and the research community have raised concerns about content moderation (e.g. [82, 119, 39, 38, 95]), including the inconsistency and unfairness of decisions, censorship, as well as the harm moderators experience as part of their work [115]. These concerns highlight the need to empower users with greater authority over their comment sections—both as creators and viewers. Rather than relying solely on opaque, automated systems, users should have more control over which comments they see and how they are organized. By enabling users to cluster comments based on their priorities—such as sentiment, topic, or relevance—this approach not only improves their experience but also strengthens their agency in navigating online spaces. My thesis explores how such user-centered moderation tools can be designed to balance freedom of expression with digital well-being.

1.1 The Motivation Behind Dealing with Hate Comments

The following examples highlight the impact of hate comments, motivating my work on this topic. While some individuals channel negativity into positive outcomes, many struggle to cope with such hostility.

Madylin Bailey transformed hate comments from YouTube into inspiration, composing a song she performed on America's Got Talent in 2021. Her audition went viral, catapulting her to fame overnight [5].

In contrast, Rainer Winkler, known as "Drachenlord", faced tragic consequences due to his engagement with hate comments. His response videos and call-outs escalated when he revealed his home address, leading to harassment and a physical confrontation with a "hater". This resulted in a one-year probation sentence for Rainer Winkler, narrowly avoiding a two-year prison term [110, 122, 31, 111].

[3] The Chrome extension "Tune" can be used to filter comments based on their toxicity level on social media platforms but does not allow users to define their own criteria `https://github.com/conversationai/perspective-viewership-extension`

Joshua Weissman, a popular YouTube chef, addressed hate comments in a video, revealing that after losing 60 lbs, criticism shifted from calling him overweight to claiming he was too skinny to be a good chef [63]. This demonstrates how negativity persists regardless of personal change.

1.2 The Challenge of Comment Moderation

Research on sentiment analysis (discussed in greater detail in Chapter: 2) highlights YouTube's reliance on machine learning for comment moderation. However, these algorithms often fail to detect hate speech, toxic comments, and other harmful content due to the complexity, context, and evolving nature of language [48]. While human moderation exists, it primarily focuses on video content at upload. For comments, YouTube's *Priority Flagger Program* [124] enables government agencies and NGOs to report guideline violations, which are then prioritized for review within 24 hours by YouTube's Trust & Safety Team [56].

Initiatives like Google Jigsaw's *Perspective API* [90] have faced academic critique for technical limitations and biases in detecting *toxicity* [48]. For instance, Caroline Sinders' 2017 post [105] revealed that the word "arabs" was initially flagged as 63% toxic, though this has since dropped to 11.27%. Yet, the API still struggles with context, classifying e.g., "Ozempic face" [4] as only 16.76% toxic and 8.36% insulting.

These limitations expose creators and consumers to harmful comments, degrading their online experience. To address these challenges, this work proposes an alternative approach using Small Language Models (SLMs), which offer efficiency, adaptability, and lower computational costs than Large Language Models (LLMs). In the following sections, the overall aim of this work and the specific research questions will be outlined. Additionally, an analysis of existing language models will be conducted, exploring the potential of SLMs for real-time, user-centered moderation. Finally, the rationale behind selecting specific language models for this research will be explained.

1.3 Aim of Work

This thesis develops a browser extension for sentiment analysis of YouTube comments using Small Language Models (SLMs). The project encompasses a comparison of 20 on-device language models—DeepSeek-R1 1.5b, 7b, 8b, Gemma2 2b, 9b, InternLM2 1.8b, 7b, Llama 3.2 1B, 3b, mistral-small 22b, Phi-3.5 3.8b, Phi-4 14b, Qwen2.5 0.5b, 1.5b, 3b, 7b, SmolLM 135m, 360m, 1.7b, and a fine-tuned BERT model, DistilBERT-sst2—evaluating their accuracy, performance metrics, and false positive/negative rates compared to each other and a human baseline for a small dataset of YouTube comments. In parallel, a

[4] The growing popularity of diabetes medications like Ozempic® for weight loss has led to increased use among non-diabetic patients. This trend has sparked the term "Ozempic face" on social media, referring to facial changes—such as extreme weight loss, distorted contours, and skin sagging—observed in both diabetic and non-diabetic users [3]

human-centered approach investigates user preferences and the effectiveness of various methods for presenting positive, neutral, and negative comments and also experiments with a multi-tier classification system within the YouTube comment section.

Three tools (Tool A, Tool B, Tool C) are developed and evaluated through user testing, each with distinct features for sentiment analysis and comment moderation. The characteristics and functionalities of these tools will be detailed in the next section, as well as in Section 5.3. Through an iterative design process involving user interviews and prototype evaluations, the research guides the development of the extension. The main part of this work is the creation and testing of a browser extension, leveraging the most proficient on-device language model and adhering to user-preferred display methods. Finally, a comprehensive user study assesses the extension's impact on user experience with classified YouTube comments, providing insights into the efficacy of on-device sentiment analysis using SLMs in real-world social media contexts. This leads to the following research questions and hypotheses, described in the next section.

1.4 Research Questions & Hypotheses

The research questions and hypotheses outlined below address distinct aspects of this thesis, spanning technical, human-centered, and bridging technical and human-centered perspectives. All research questions and hypotheses are designed to explore a specific dimension of sentiment analysis and user interaction with YouTube comments.

RQ1: **(Technical)** How do the accuracy, efficiency, and false positive/negative rates of general-purpose on-device Small Language Models compare to a specialized, fine-tuned BERT-based language model in doing sentiment analysis of YouTube comments?

HYP1: SLMs will perform better in terms of identifying positive, negative or neutral comments on YouTube than a specialized, fine-tuned BERT-based language model.

RQ2: **(Human-centered)** What are user preferences and the perceived effectiveness of different methods for displaying (grouped, reordered, highlighted) potentially negative YouTube comments in a browser extension, considering the ethical implications of false positives and negatives?

HYP2: Users will prefer a comment with positive sentiment hidden instead of a comment with negative sentiment displayed due to a false positive in the sentiment analysis.

RQ3: **(Bridging technical and human-centered)** How does the use of Small Language Models compare to a state-of-the-art BERT-based model and influence users' perceived usefulness of sentiment analysis systems for YouTube comments?

HYP3: Users perceive SLMs as more useful, due to their contextual understanding and capability to explain what contributed to their classification of a comment.

These research questions and hypotheses are used to guide the evaluation of language models and user interaction methods in this thesis. In the next section, an overview of the models best suited to address these questions will be given, and their suitability will be explained, laying the groundwork for the subsequent user study and prototype development. A detailed analysis will be presented in Chapter: 4.

The overall structure of this thesis begins with an overview of related work, followed by a detailed explanation of the applied methods. Next, a chapter presents an in-depth evaluation of the models and related findings. This is followed by a dedicated chapter on the results of the user evaluation (see Chapter: 5). Finally, the thesis concludes with a general discussion of the findings and the conclusion.

1.5 Overview: Language Models

In the following, the differences between BERT (Bidirectional Encoder Representations from Transformers) and SLMs in the context of sentiment analysis are explored, focusing on the hypothesis that SLMs perform better in this domain. The motivation for this comparison comes from preliminary tests (see Section 1.5.2) which further resulted in the development of three distinct tools during this thesis for sentiment analysis of YouTube comments. Tool A uses DistilBERT-sst2 as a state-of-the-art, BERT-based baseline, while Tools B and C leverage an SLM with unique approaches, including video transcript context integration, reasoning capabilities, three-way classification (Tool B), and a multi-tier classification system (Tool C). By examining the functional distinctions between BERT and SLMs, this chapter will lay the groundwork for evaluating their performance in the subsequent model evaluation and user research.

1.5.1 What is BERT?

BERT is a transformer-based, encoder-only model that has become a benchmark in natural language processing (NLP) tasks since its introduction by Google in 2018 [27]. It employs a masked language modeling (MLM) objective, where a portion of the input tokens are masked, and the model learns to predict them using bidirectional context. This bidirectional understanding allows BERT to capture rich semantic relationships within the text, making it strong in tasks such as classification and question-answering. However, despite its strengths, BERT presents certain limitations in the realm of sentiment analysis. Although it performs well on general datasets, it struggles with domain-specific nuances such as the informal and context-dependent nature of YouTube comments. Additionally, BERT outputs confidence scores without contextual reasoning, which can limit the interpretability of the results. This limitation is particularly critical in sentiment analysis, where understanding sentiment drivers is important, as well as in identifying potential biases built into the model. The need for interpretability is emphasized by the warnings on the BERT model page, which highlight potential bias in the model training data [26].

DistilBERT Fine-Tuned on SST-2 for Sentiment Analysis

For this study, Tool A utilizes DistilBERT-sst2 [54], a smaller and faster variant of BERT that retains much of its performance while reducing computational requirements. DistilBERT is trained using knowledge distillation, where a smaller model (the student) learns from the outputs of a larger, fully trained BERT model (the teacher). To specialize DistilBERT for sentiment analysis, it was fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset [107]. SST-2 is a dedicated dataset for binary sentiment classification, where each sentence is labeled as either positive or negative. While fine-tuning on this dataset enhances DistilBERTs sentiment analysis capabilities, it also introduces the challenge of adapting a model pre-trained on rather general sentiment data to a specific domain like YouTube comments. The models reliance on confidence scores rather than contextual reasoning further highlights the need to evaluate whether SLMs might offer better performance in delivering nuanced and interpretable sentiment analysis results.

1.5.2 Advantages of Small Language Models

The field of artificial intelligence (AI) has seen remarkable progress, particularly in natural language processing and understanding. The release of ChatGPT in November 2022 sparked widespread interest in LLMs, showcasing their impressive capabilities in understanding and generating human-like text [84]. However, the high hardware requirements for running LLMs have led to the emergence of an alternative trend: the rise of Small Language Models (SLMs), e.g., Gemma 2 2b, Llama 3.2 1b, Qwen2.5 0.5b. These compact models are designed for simpler tasks and are gaining traction due to their lower computational demands and recent advancements enabling them to operate directly within web browsers using tools like WebLLM [74].

In preliminary tests, I compared three approaches for sentiment analysis (POSITIVE, NEGATIVE, NEUTRAL) using the phrase "Ozempic face":

- A traditional TextBlob approach inspired from the literature [68], which classified the phrase as *NEUTRAL* (0.0).
- A BERT-based model (*distilbert-base-uncased-finetuned-sst-2-english*), which classified it as *POSITIVE* (0.5789).
- A Generative Pre-trained Transformer (GPT) approach using Google's Gemma 2 2b, which labeled it as *NEGATIVE* and provided the following reasoning: *"The term 'Ozempic face' is often used to describe a side effect of the medication, implying a negative cosmetic outcome."*

These results reinforced my assumption in the potential of SLMs for sentiment analysis, particularly due to their ability to provide context-aware reasoning and accurate classifications.

The shift toward efficient, localized AI is further supported by Google's proposal for a web API for prompting language models within the Chrome browser, highlighting the growing accessibility and significance of these technologies [36]. On-device language models, in particular, could offer a unique solution to content moderation challenges. By enabling real-time, context-aware analysis directly within the user's browser, they address privacy concerns, reduce reliance on cloud processing, and empower users with greater control over their online experience.

SLMs are characterized by their relatively low parameter count compared to LLMs. While they may lack the extensive pre-training and capacity of larger models, SLMs excel and even outperform larger models in specialized tasks such as sentiment analysis when fine-tuned [125]. Their compact architecture enables faster inference, reduced resource consumption, and greater flexibility for fine-tuning. In this study, Tools B and C demonstrate how SLMs can perform sentiment analysis even without being explicitly fine-tuned, soley by prompting. Additionally, Tool B and C will incorporate a video transcript as context to better understand and guide the classification of each comment. This way, SLMs can also offer interpretable explanations and deeper insights for classification decisions—addressing a key limitation of BERT-based models further detailed in Section: 4.6. Tool C introduces a multi-tier classification approach, categorizing comments into 20 distinct sentiment classes (see Appendix: 7), identifying primary classes (e.g., Praise/Appreciation, Constructive Criticism), tones (e.g., sarcastic, neutral), and flagging specific content types (e.g., hate speech, political/religious references). Unlike BERT-based models, which require extensive fine-tuning for specific tasks, SLMs operate on prompts, leveraging their built-in world knowledge to adapt to new scenarios without additional training. This flexibility, combined with local deployment via Ollama [83], ensures data privacy and makes SLMs highly suitable for real-world applications.

1.5.3 Key Differences: BERT vs. SLMs in Sentiment Analysis

When applied to sentiment analysis, BERT and SLMs demonstrate distinct strengths and limitations. BERT, with its robust bidirectional context understanding, generally achieves high accuracy in standard sentiment classification tasks. Its transformer-based architecture allows it to grasp the sentiment of a sentence by considering the surrounding text, up to a certain point. However, BERTs approach can be less effective in scenarios that require contextual or nuanced sentiment interpretation, such as analyzing YouTube comments where sentiment might be influenced by e.g., external events, specific video content or community-specific language.

In contrast, SLMs, as demonstrated by Tools B and C in this study, offer a more adaptable approach to sentiment analysis. By leveraging the video transcript as additional context, these tools enhance the model's ability to understand the sentiment behind comments in relation to the video content. They also provide explanations for their classification decisions, contributing to greater transparency and interpretability. Tool C further showcases the potential of SLMs by implementing a multi-tier classification system, proving that the same model can achieve more granular sentiment analysis without the need for prior fine-tuning for specific tasks.

Notably, most SLMs inherit multilingual capabilities and the ability to interpret mixed-language from their base model, as well as emoji interpretation out of the box. These capabilities highlight how general purpose SLMs, could outperform specialized models like DistilBERT-sst2 in specific, context-rich applications, particularly when adaptability and interpretability are critical success factors.

The sections above explored the differences between BERT and SLMs in sentiment analysis, posing the hypothesis that SLMs outperform BERT-based models in this domain. To situate this work within the broader research landscape, the next section explores existing advancements in sentiment analysis, hate speech detection, community learning and value-sensitive algorithms, text classification with language models, content moderation, personalized filtering, human-in-the-loop systems, and ethical considerations. It highlights key challenges, gaps, and related research areas that this work builds upon or draws inspiration from.

CHAPTER 2

Related Work

The rapid growth of social media platforms has introduced significant challenges in managing online discourse, with issues such as online hate and microaggressions becoming increasingly prevalent [8]. Automating the detection and classification of such content is crucial for fostering healthier online environments.

While numerous methods have been proposed to address these challenges, their effectiveness remains limited. A key issue is the highly contextual nature of hate speech, which complicates accurate interpretation and classification for existing approaches [50].

The following sections explore relevant research areas, including sentiment analysis, hate speech detection through prompting, community learning and value-sensitive algorithms, text classification with language models, interactive content moderation, personalized filtering, human-in-the-loop systems, and fairness in content classification. These areas collectively form the foundation for the proposed solution, which seeks to empower users with greater control over the content they encounter in YouTube comments.

2.1 Sentiment Analysis

Sentiment analysis, a key area in affective computing, involves detecting, analyzing, and understanding peoples attitudes, opinions, and emotions toward an entity, like events, services, or other topics [65]. Sentiment refers to a thought, attitude, or judgment influenced by emotions, making it distinct yet closely related to opinions. For example, emotions often drive people to form judgments and opinions about something. While sentiment and opinion are closely linked, they are not the same—sentiment reflects an emotional perspective, while opinion is a more reasoned judgment [65]. Building on this foundation, practical applications of sentiment analysis have emerged, with Google Jigsaws “Perspective API” standing out as a notable initiative.

The "Perspective API" leverages machine learning to enhance online discussions by classifying text into attributes such as *Toxicity*, *Severe Toxicity*, *Identity Attack*, *Insult*, *Profanity*, and *Threat* [90, 89]. Since its launch in 2017, it has faced academic critique for technical limitations and biases, prompting the Conversation-AI team to address these issues through de-biasing efforts [48, 105, 60, 59].

The "Perspective API" has been widely adopted, with more than 1,000 partners, including major platforms such as Reddit and publications like the New York Times, processing nearly two billion requests per day in 18 languages [61]. It has also been integrated in datasets like "RealToxicityPrompts" to measure toxicity in outputs from LLMs such as LLaMA, Flan-PaLM, and InstructGPT [42, 114, 20, 86]. Additionally, research shows that up-ranking constructive or curious comments can improve prosocial behavior in online discussions [96]. Sentiment analysis on platforms like YouTube has revealed that positive comment sentiment correlates with longer video watch times, while negative sentiment shortens engagement [123]. However, challenges remain, such as multilingual comments, context awareness, and capturing nuanced sentiment. Recent meta-reviews highlight the limitations of traditional approaches (e.g., decision trees, SVMs) and deep learning methods, noting that research often focuses on specific languages or content genres without generalizability [77].

2.1.1 Multimodal Sentiment Analysis

Introduced nearly two decades ago, sentiment analysis has evolved from extracting attitudes and opinions from text to a powerful tool widely used in government, business, medicine, and marketing. With technological advancements, it has expanded into multimodal approaches, incorporating audio, image, and video, significantly enhancing its effectiveness [24]. Recent research highlights substantial progress in multimodal sentiment analysis, underscoring its growing importance and potential [108]. This evolution makes it particularly well-suited for platforms like YouTube, where diverse content formats enable more comprehensive and accurate sentiment detection [24].

Early work in 2013 demonstrated the potential of multimodal sentiment analysis by integrating text, audio, and visual features to analyze movie review videos. This approach combined speech-based emotion recognition (audio) and valence information from video, showing that training on written movie reviews could be as effective as using spoken in-domain data for analyzing spoken review videos [120].

2.1.2 Irony, Sarcasm, Emoji Detection

Irony and sarcasm, where speakers imply the opposite of their intended meaning, are complex linguistic phenomena that pose significant challenges for computational detection due to their reliance on social and cognitive nuances. Irony often highlights unexpected events, while sarcasm conveys criticism, complicating tasks like sentiment analysis [93, 37]. Early work identified surface patterns, such as punctuation and emoticons, in ironic sentences, but sarcasm detection remains challenging. Lexical factors, like parts of speech and punctuation, have been shown to aid in identifying sarcasm, even when explicit cues are removed [47, 37].

Emojis, increasingly used in digital communication, play a significant role in sentiment analysis. While congruent emojis amplify sentiment, incongruent ones may signal sarcasm or ambiguity [94, 112]. Subramanian et al. (2019) improved sarcasm detection by integrating emojis into neural networks, though challenges related to data and context remained [112]. Emojis have also enhanced emotion detection in complex linguistic contexts. For instance, Liu et al. (2021) developed CEmo-LSTM, combining emojis with Bi-LSTM, to analyze Chinese Weibo data during COVID-19, reaching high accuracy (~0.95) and revealing a rise in negative emotions [71]. Similarly, Rendalkar et al. (2018) used emojis and lexical resources like WordNet to classify social media posts into eight emotions, providing a nuanced understanding of sentiment and sarcasm [92].

This thesis leverages the language-agnostic and general-purpose capabilities of SLMs to address challenges like Emoji and mixed language detection as well as contextual awareness offering a novel approach to sentiment analysis. The identified limitations, such as multilingual nuances and context awareness, are discussed in the methodological approach (see Chapter: 3).

2.2 Hate Speech Detection Through Prompting

The rise of AI-based models has revolutionized hate speech detection by enabling systems to capture complex contextual nuances in text. Unlike traditional fine-tuning, prompting has emerged as a more efficient approach, requiring less data and computational resources while achieving competitive results [8]. Recent advances demonstrate that LLMs outperform conventional methods in identifying hate speech, particularly when optimized prompting strategies are employed. These findings underscore the critical role of well-designed prompts in maximizing the effectiveness of LLMs for accurate and context-aware hate speech detection [50].

While these studies focus on hate speech, my work applies similar principles to classify YouTube comments with SLMs additionally providing explanations on classification results for transparency and enabling users to navigate content more consciously.

2.3 Community Learning and Value-Sensitive Algorithms

Automated moderation systems are increasingly incorporating socio-technical approaches, such as community learning and value-sensitive algorithms, to improve decision-making. By leveraging past moderator actions and integrating stakeholder knowledge, these methods enhance the accuracy and relevance of moderation tools. For instance, cross-community learning has been successfully applied to develop moderation systems for platforms like Reddit, demonstrating the value of knowledge transfer between diverse communities [19]. Such approaches highlight the importance of aligning algorithmic systems with human values and community-specific insights to create more effective and equitable moderation frameworks [41, 127, 106].

Inspired by these principles, my research focuses on designing a user-centric browser extension for YouTube that empowers users to filter and interact with comments based on sentiment, rather than relying on top-down moderation or predefined rules.

2.4 Text Classification with Language Models

Text classification has been a cornerstone of natural language processing for a long time, with language models playing a pivotal role in tasks such as topic detection, spam filtering, and sentiment analysis [10]. While neural networks have set new benchmarks in NLP, their resource-intensive training requirements have led to the widespread adoption of pre-trained models like BERT, which offer state-of-the-art performance with reduced computational costs. However, the generic nature of these models often limits their effectiveness in domain-specific applications. Surprisingly, even domain-specific models like FinBERT, designed for financial texts, do not consistently outperform generic models like RoBERTa, suggesting that domain specialization alone may not address the challenges of nuanced classification tasks [7].

The focus on English in hate speech detection research has left many low-resource languages, such as Arabic and Ewe (an African language that is highly underrepresented, linguistically complex, and lacks standardized datasets), underexplored [1, 9]. Arabic's linguistic complexity complicates the use of generic models, while Ewe faces challenges due to poor-quality, unbalanced datasets [9]. Recent studies show that language-specific approaches, such as Arabic pre-trained models (e.g., DziriBERT, AraBERT v2) combined with data augmentation, significantly improve performance. Similarly, for Ewe, researchers have developed a high-quality dataset and fine-tuned transformer-based models, with BERT-base-cased achieving exceptional accuracy (0.972) and F1-score (0.970), setting a new benchmark for low-resource NLP tasks [1]. These results highlight the importance of language-specific models and tailored datasets for addressing the challenges of low-resource languages.

While my work does not focus on low-resource languages, it similarly addresses the challenge of adapting models to the informal and varying language of YouTube comments. By comparing SLMs with a BERT-based approach, I aim to identify a solution that balances performance, transparency, and user preference, ensuring that users can navigate comments more effectively in a user-friendly way.

2.5 Content Moderation

Recent HCI and social computing research has focused increasingly on content moderation, examining its role in fostering online communities, enhancing engagement, ensuring safety, and addressing societal concerns [58]. Content labels, such as fact-checks and sensitive content warnings, have emerged as tools to counter disinformation. However, moderation faces challenges like automated bots and varied user motivations for sharing misinformation, from humor to genuine belief. Content labels provide a flexible alternative to binary moderation (removal or downranking), balancing user autonomy with effective misinformation mitigation [78]. Studies also highlight how user values—such as diversity, inclusion, and compassion shape demands for culturally competent moderation, transparency, and equitable tools, urging platforms to prioritize compassion over profit. This underscores the need for relational and restorative justice to maintain community connections and repair harm, particularly during contentious moderation decisions [115]. A recent trade-off-centered framework further advances this field by exploring interconnected layers of moderation—actions, styles, philosophies, and values—emphasizing the need to weigh trade-offs and consider their broader implications for communities [58].

While these works focus on moderation, my research shifts to user empowerment through sentiment analysis and personalized filtering. My browser extension uses SLMs to provide sentiment labels (positive, neutral, negative) and filtering options, enabling users to navigate YouTube comments more consciously and transparently.

2.6 Personalized Filtering

Already in 1993, research explored how techniques from "Artificial Life", specifically a combination of "Genetic Algorithms" and "Learning from Feedback", could be used to create personalized information filtering systems. These systems dynamically adapted to users' changing interests, ensuring the filtering process remained relevant over time. A prototype was developed to assist users in retrieving articles, with experiments showing that the genetic algorithm improved the recall rate (capturing more relevant content), while the learning mechanism enhanced the precision rate (reducing irrelevant results) [99].

Recent studies on personalized filtering in social networking sites highlight the "filter bubble" problem, where users are only shown content matching their preferences, limiting exposure to new topics. To address this, researchers proposed interactive visualizations to increase transparency and user control, improving trust and experience. Evaluations showed that the visualization helped users to better understand filtering mechanisms and that it not only increased users' trust in the system but that they also felt more in control [81].

Another study on mobile adaptive personalization systems explored whether adaptive personalization improves over time, outperforms self-customization, and benefits from social network data. A mobile news personalization system was developed, adapting content based on user behavior and social networks. Findings revealed that: (1) adaptive personalization improves over time, (2) it outperforms self-customization, and (3) incorporating social network data further enhances personalization. These results suggest that adaptive systems leveraging social networks are highly effective [21].

2.7 Human-in-the-Loop Systems

Understanding how individuals interact with—and prefer to interact with—machine learning systems is essential for designing effective and user-friendly systems [6]. Terms like "human-AI collaboration", "human-AI partnership", and "human-AI teaming" highlight a shift away from full automation, emphasizing the importance of effective collaboration between humans and AI. Such collaboration is critical for safety in high-stakes domains, leveraging the complementary strengths of humans and machines to achieve better outcomes, reduce computational complexity, and enable innovative technologies beyond current AI capabilities. A common approach is "human-in-the-loop", where human input enhances AI performance, as seen in interactive Machine Learning (IML) [69]. IML integrates human insight and domain knowledge with the computational power of machines, actively involving users while reducing the need for specialized expertise typically required in traditional machine learning methods [28, 69]. Importantly, research shows that the success of IML heavily relies on well-designed interfaces that facilitate intuitive and seamless interaction between humans and machines [28]. Research on recidivism prediction has compared human judgment to algorithmic performance, with mixed findings. While humans and algorithms can perform similarly under controlled conditions, algorithms often outperform humans in realistic, complex scenarios, particularly when using enriched risk factors or when humans lack feedback. These results highlight the importance of context in evaluating human and algorithmic decision-making for predictions [70].

While human-in-the-loop systems often target high-stakes domains or expert-driven tasks, my work aims to improve trustworthiness and empower users in everyday contexts, specifically for navigating YouTube comments. Rather than relying on full automation or complex systems, my browser extension prioritizes transparency and user control, enabling individuals to interact with sentiment classifications in an informed way that aligns with their preferences. By leveraging SLMs that provide interpretable reasoning, my approach builds on "human-in-the-loop" principles, helping users make informed decisions about the content they engage with.

2.8 Fairness in Content Classification/Ethical considerations

As AI performance continues to advance, often surpassing human capabilities on benchmark datasets [32, 117], AI models show significant potential to enhance human decision-making across a wide range of domains [121, 51, 64]. However, full automation is not always desirable due to ethical, legal, and safety concerns, particularly in high-stakes applications. This has led to a growing emphasis on human-AI collaboration, where human oversight and input are integrated to ensure responsible and effective use of AI systems [69]. Algorithmic filtering can have biases, both technical and human, which significantly impact online services. Research shows that these biases can arise from societal prejudices, technical limitations, or emerge after deployment, leading to ethical concerns such as reduced informational diversity, compromised user autonomy, and limited transparency. Research emphasizes that algorithmic gatekeeping, which combines human editors and machine code, does not eliminate human biases. Instead, they highlight the need to design systems that promote values like autonomy, transparency, and equitable access to information [11]. A recent study explored how YouTubers perceive the fairness of YouTubes content moderation, highlighting the importance of equal treatment, consistent decisions, and having a voice in algorithmic processes. The study found that creators view moderation as unfair when treated unequally, faced with inconsistent decisions, or excluded from influencing algorithms. It emphasizes the need for transparency, such as disclosing how moderation impacts revenue and visibility, and incorporating creator input into decision-making [72].

My work complements these findings, but shifts the focus to YouTube users rather than creators. By providing tools with sentiment analysis and filtering options, I aim to empower users to navigate comments more effectively, prioritizing transparency and user control. This approach aligns with the prior referenced research on fairness and accountability while addressing the needs of everyday users.

CHAPTER 3

Methodological Approach

In this thesis, a mixed-methods approach [22] is employed to evaluate the technical performance of SLMs as well as the user needs for a browser extension designed to perform automated sentiment analysis of YouTube comments. This approach is chosen to ensure that both the technical robustness of the models and the user-centric design of the browser extension are addressed.

The **quantitative** approach consists of the evaluation of DistilBERT-sst2 and 19 SLMs against a human baseline in terms of binary (positive/negative) and three-way (positive, neutral, negative) sentiment analysis. Macro-F1 scores (see Section: 4.3.1) and performance metrics (see Section: 4.3.2) are used to provide comparable benchmarks for model performance. The macro-F1 score is selected because it balance precision and recall, making them suitable for evaluating sentiment analysis tasks with multiple classes with a small sample size. The dataset used for evaluation is described in Section 4.4.

The **qualitative** approach consists of two main studies: a user preference study (see Section: 5.1) evaluating prototypes in an interview setting and a user evaluation study (see Section: 5.4) for evaluating the final tools. In the user preference study, semi-structured interviews, think-aloud [30] protocols, and note-taking are conducted to gain insights into user preferences, experiences, and concerns regarding the prototypes for automated sentiment analysis.

The user evaluation study builds on the findings from the user preference study and includes a pre-task and a post-task questionnaire, a task sheet with exercises, and a semi-structured interview. Participants are asked to think aloud [30] during the audio-recorded tasks and interviews, while notes are taken by the researcher. The pre/post-task questionnaires are designed to include both Likert-scale [62] and open-ended questions to assess changes in user experience and perception of online discourse. All materials can be found in the Appendix: 7. A thematic analysis as proposed by Braun and Clarke [12, 13] is applied to the interview transcripts to identify patterns and themes related to user needs and ethical considerations and to provide a deeper understanding of user interactions with the browser extensions.

Before the user studies are conducted, the necessary *ethical approvals* [1] are obtained. Throughout the research process, participant privacy and data protection are prioritized to ensure compliance with ethical standards. Additionally, potential biases inherent in language models are addressed, and their limitations in accurately performing sentiment analysis are acknowledged. Beyond technical aspects, the broader societal implications of deploying automated content moderation technologies are also considered.

To summarize, the research is structured into four main phases:

- **Phase 1: Language Model Evaluation** – A quantitative evaluation of SLMs and BERT-based models is conducted to identify the best-performing models.
- **Phase 2: Prototype Development & User Preference Study** – After developing prototypes for displaying and interacting with classified comments, a qualitative study is carried out to understand user needs and preferences for the browser extension.
- **Phase 3: Browser Extension Development** – The Browser extension is designed and implemented based on findings from Phases 1 and 2.
- **Phase 4: User Evaluation Study** – A mixed methods study is conducted to evaluate the Browsers' extensions usability, trustworthiness and effectiveness. The study is recorded and transcribed to enable a thematic analysis of the qualitative data.

It is important to note that the best-performing model for the tasks in this thesis can only be determined after the user evaluation study (Phase 4), as the participants will serve as annotators to establish the ground truth. Therefore, the model used in the user evaluation study is based on an educated assumption derived from prior tests conducted during Phase 1.

The following sections describe the methodological approach for each phase of the study. First, **Phase 1: Language Model Evaluation** is outlined, detailing the preparation of the dataset and model selection criteria. Second, **Phase 2: Prototype Development and User Preference Study** is described, focusing on the creation of prototypes and the study design, which includes semi-structured interviews. This study design is also applied in Phase 4, as both phases share methodological similarities in their use of interviews. Third, **Phase 3: Browser Extension Development** is covered, explaining the differences in design choices and implementation methods based on findings from Phase 1 and 2. Finally, **Phase 4: User Evaluation Study** is discussed, with a focus on

[1] `https://www.tuwien.at/en/research/rti-support/responsible-research-practices` - last accessed 19.03.2025

the pre-task and post-task surveys, training video and user tasks as well as the thematic analysis of the interview data.

3.1 Phase 1: Language Model Evaluation

In this section, the methodological steps for evaluating language models are outlined. First, a balanced dataset of YouTube comments with neutral, positive, and negative content is prepared. Second, SLMs and a BERT-based model are selected and assessed for sentiment analysis, with their performance informing an educated assumption for Phase 4. Third, a comparative analysis is conducted using a dataset annotated by humans in Phase 4, with metrics like macro-F1 score calculated. Finally, the results are analyzed to address Research Question 1 (RQ1): "How do the accuracy, efficiency, and false positive/negative rates of general-purpose on-device Small Language Models compare to a specialized, fine-tuned BERT-based language model in doing sentiment analysis of YouTube comments?". While this section focuses on the methodological approach, detailed results are discussed in Section 4.

The growing number of SLMs necessitated a systematic evaluation to determine which model to use for the user evaluation study. The selected models were evaluated for their performance in traditional sentiment analysis (positive, negative, neutral) and for multi-tier classification while also using them to explain their classification results and to set them into context.

Initial tests used the IMDb dataset [73], where SLMs were prompted to perform binary sentiment classification and provide reasoning for their assessments. A random sample of 20 comments from the 12,500-comment IMDb training corpus, which exclusively contained negative comments, was used for this analysis. A Streamlit [2] application was developed to facilitate the evaluation (see Appendix: 7). Interestingly, inconsistencies emerged: the SLM (Phi-4:14b) classified 4 out of 20 comments as positive, contradicting the dataset's negative labels.

To address the limitations of binary classification, a follow-up test introduced a three-way classification system (positive, negative, neutral) using the same 20 comments. This approach yielded more nuanced results, with 2 of the previously "misclassified" comments identified as positive and 3 (including two previously classified as positive) as neutral. The full analysis, including the comments, classification results, and detailed reasoning provided by the SLM, is presented in Appendix: 7.

The models that were compared based the dedicated YouTube dataset (see Section: 4.4), along with their key features and performance, are analyzed in detail in Chapter 4, which provides a comprehensive evaluation of their strengths and limitations. In the following section, the development and testing of the prototypes are described, along with the study design.

[2] https://streamlit.io/ - last accessed 23.03.2025

3.2 Phase 2: Prototype Development & User Preference Study

This section explains the approach for developing the prototypes and the study design for Phase 2, which, due to its methodological similarities with Phase 4, also applies to the latter. First, the development and testing of prototypes for displaying and interacting with classified comments are described. These prototypes are used for the interviews during the user preference study. Second, Section 3.2.3 details the structure of the interviews conducted in both Phase 2 and Phase 4, highlighting their shared approach. The process of collecting qualitative data through semi-structured interviews with YouTube users is explained (the evaluation for both phases is described in Section 5). Finally, the results are analyzed to address Research Question 2 (RQ2): "What are user preferences and the perceived effectiveness of different methods for displaying (grouped, reordered, highlighted) potentially negative YouTube comments in a browser extension, considering the ethical implications of false positives and negatives?". While this section focuses on the methodological approach, detailed results are discussed in Chapter 5.

3.2.1 Prototype Development

The development of the first prototypes followed an iterative process, starting with an initial concept and evolving into five interactive Figma [3] prototypes (further details on the prototypes can be found in Chapter 4). Both the prototypes and the final browser extension were designed with the goal of analyzing sentiment in YouTube comments while prioritizing user preferences for comment display and interaction methods, ensuring user privacy through local model hosting, and providing reasoning for the model's decisions where possible. The prototypes, findings from the user preference study, and the rationale behind design choices are discussed in greater detail in Chapter 5.

3.2.2 Participant Recruitment: User Preference Study

After designing the prototypes, participants for the user evaluation study were recruited. The study employed a convenience sampling approach [35], reaching out to 10 potential participants (6 male, 4 female) via social media channels and private networks. Participation in the study required proficiency in English and regular use of YouTube, while a technical background was not necessary. To ensure a broader range of perspectives, the selection process aimed for diversity in educational degrees. Ultimately, 6 YouTube users participated in semi-structured interviews as part of the user preference study (see Table 3.1).

[3] https://www.figma.com/ - last accessed 23.03.2025

Participant	Age	Highest level of education	YouTube usage
P1	30	High school diploma	Daily
P2	30	Master of Science	Daily
P3	32	High school diploma	Daily
P4	29	High school diploma	Daily
P5	19	Compulsory school diploma	Daily
P6	34	Master of Arts	Daily

Table 3.1: Overview of participants from the user preference study, their highest level of education and YouTube usage.

3.2.3 Interview Design

In both Phase 2 (user preference study for the prototypes) and Phase 4 (user evaluation study for the final tools), semi-structured interviews were conducted. The primary difference between the phases lies in the context and setting: Phase 2 interviews were conducted remotely via Google Meet [4], while Phase 4 interviews took place in person. All interviews were audio-recorded using Open Broadcaster Software (OBS) [5], and notes were taken in a prepared document containing the interview questions. The interview questions for Phase 2 are provided in Appendix: 7, and those for Phase 4 are in Appendix: 7.

The transcription process for the interviews is described below, along with how the transcripts were used.

The audio recordings from both phases were processed using a custom-built Gradio [6] web interface (see Appendix: 7) that facilitated interaction with OpenAI's Whisper [85] large-v3 transcription model running locally on the researcher's device. Within the software, the audio files were divided into five-minute segments to improve the accuracy of the transcription. The output consisted of segmented audio files, each with corresponding text files containing time-stamped transcriptions.

For the user preference study (Phase 2), the transcripts were combined with the interview notes to identify key themes and trends that guided the development of the browser extension.

For the user evaluation study (Phase 4), a separate document was created for each interview. The AI-generated transcript was manually reviewed and corrected by comparing it with the original audio. Once verified, time codes were removed, and the text was formatted so that each interview question was followed by the corresponding participant response. Filler words (e.g., *"ehm," "mhm," "ah," "okay"*) and irrelevant interruptions (e.g., *"Oh, my cell phone is falling out of my pocket [Oh, mein Handy fällt aus der Hosentasche]"*, *"We are stopping the recording now [Wir stoppen die Aufnahme jetzt]"*, or third-party interruptions like *"May I interrupt for a second? [Kann ich eine Sekunde*

[4] https://meet.google.com/ - last accessed 23.03.2025
[5] https://obsproject.com/ - last accessed 23.03.2025
[6] https://www.gradio.app/ - last accessed 23.03.2025

stören?]") were omitted. Responses were structured into cohesive paragraphs corresponding to each question and transferred into an Excel sheet, which served as the basis for coding during the thematic analysis for Phase 4, further described in Section 3.4.8.

The next section provides a detailed discussion of the browser extension development, focusing on the creation of the three tools (Tool A, Tool B, Tool C) that were compared later in Phase 4. It describes their architectural differences, and the unique features that set them apart.

3.3 Phase 3: Browser Extension Development

In this section, the methodological steps for developing the browser extension are outlined. First, the browser extension is designed based on findings from Phases 1 and 2, incorporating insights from the language model evaluation and user preferences study. Second, the extension is implemented using a SLM selected based on the researchers assumptions and the preferred display method identified in Phase 2. Finally, the extension is thoroughly tested to ensure reproducible results when analyzing comments, ensuring its reliability and functionality, the results are described in detail in Chapter 5.

3.3.1 From Prototype to Functional Browser Extension

The software project developed for this study employs a fully local setup, combining Ollama, a FastAPI server, and three custom-built browser extensions (Tool A, Tool B, Tool C). The goal of this architecture is to perform sentiment analysis directly within the YouTube environment while prioritizing ease of use and data privacy. The architecture for each tool is displayed in Figure 3.1.

At the core of the system is Ollama, which runs locally on the user's device and leverages SLMs for natural language processing tasks. Ollama handles model inferencing and exposes an endpoint at localhost:11434, enabling other components to request sentiment classifications through HTTP calls.

To connect the browser extensions with Ollama, a FastAPI server is used, running locally and providing three endpoints (localhost:8001, localhost:8002, and localhost:8003), each dedicated to one of the sentiment analysis tools.

Tool A provides straightforward sentiment analysis by sending each YouTube comment individually to its dedicated endpoint. The server classifies the comment using DistilBERT-sst2, returning a sentiment score (e.g., 96%), which is translated into a positive or negative label in the frontend. This tool establishes a baseline using a simple and fast approach.

Tool B incorporates the video transcript to provide contextual sentiment analysis with reasoning. The comment is sent with the transcript to the endpoint. The server prompts the SLM via Ollama to classify the comment and provide a textual explanation, which is displayed in the frontend alongside the sentiment label.

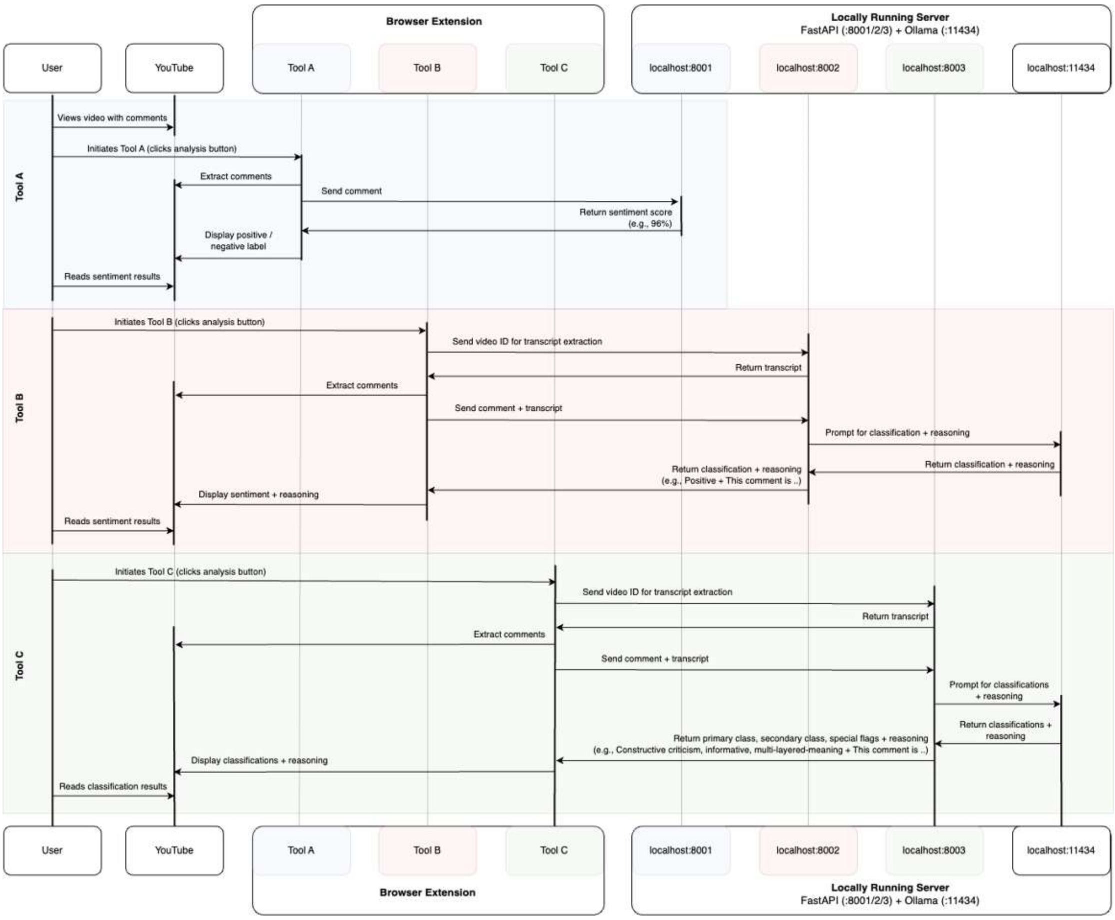

Figure 3.1: Sequence diagram illustrating the interaction flow between the user, YouTube, the browser extension, and the locally running server. The diagram details how Tools A, B, and C process YouTube comments for sentiment analysis. Tool A extracts comments and assigns a sentiment score, which is displayed as a simple positive or negative label. Tool B additionally incorporates video transcript context and provides a reasoning-based classification. Tool C further refines the analysis by applying a multi-tier classification system, categorizing comments with primary and secondary sentiment classes, special flags, and explanatory reasoning. The backend consists of a FastAPI server running locally on different ports, with Ollama handling language model inference.

Tool C employs the most sophisticated approach, performing multi-tier sentiment classification. Similar to Tool B, both the comment and video transcript are sent to its endpoint. The SLM analyzes primary sentiment (e.g., praise/appreciation, constructive criticism), tone (e.g., humorous/playful, aggressive/hostile), and special flags (e.g., hate speech, political/religious references), along with detailed reasoning for each class. A full list of Tool C's classification categories and how DeepSeek-V3 [25] was used to create them is provided in Appendix: 7.

All processing is performed locally, ensuring that no data is sent to external servers or cloud services, thereby enhancing user privacy. The scraping process only processes comments visible to the user, loading them in batches of 20 as the user scrolls. Metadata such as usernames, profile pictures, donations, likes, and replies are omitted.

A key design choice was to avoid API keys when scraping YouTube comments and transcripts directly within the extensions. This simplifies onboarding, allowing users to start immediately without configuration, thereby increasing accessibility. The extensions also feature filtering functionality, enabling users to customize which comments are displayed based on classification results.

For the user evaluation study, the initial concepts of grouping, redacting, highlighting, hiding, or reordering comments were omitted based on feedback from the user preference study. Instead, the extensions follow a non-intrusive approach, directly integrating sentiment labels and reasoning into YouTubes comment section without requiring users to navigate away from the platform. The user interface for all tools is displayed in Appendix: 7.

3.4 Phase 4: User Evaluation Study

The interview design, including the structure and methodology of the semi-structured interviews, was already described in detail in Section 3.2.3 and will not be repeated here. Instead, this section focuses on the details of how the pre-study and post-study processes were conducted, explaining the recorded training video and the user tasks, as well as how the thematic analysis was performed. Finally, the results are analyzed to address Research Question 3 (RQ3): "How does the use of Small Language Models compare to a state-of-the-art BERT-based model and influence users perceived usefulness of sentiment analysis systems for YouTube comments?". While this section focuses on the methodological approach, detailed results are discussed in Chapter 5.

3.4.1 Participant Recruitment: User Evaluation Study

The study employed a convenience sampling approach to recruit participants. The selection criteria for the user evaluation study were the same as for the user preference study: Participants required proficiency in English and had to be familiar with YouTube, while a technical background was not necessary. Similarly, a diversity in educational degrees was aimed for to cover a broader range of perspectives. Outreach was conducted via social media channels to 14 potential participants (10 male, 4 female), five of whom had previously participated in the user preference study. The remaining nine had no prior knowledge of sentiment analysis or the browser extensions being tested. Eight individuals initially agreed to participate, but one withdrew, resulting in a final sample of seven individuals, including three from the user preference study. This mix provided fresh perspectives alongside the potential for comparative insights. Participants are identified using the following notation for cross-study clarity:

- Participants from the previous user preference study are denoted with an asterisk (*)
- Cross-study participant mapping:
  - Current P3 corresponds to previous P1
  - Current P4 corresponds to previous P6
  - Current P6 corresponds to previous P3

This resulted in the following participant identifiers in the current study: P1, P2, P3*, P4*, P5, P6*, P7

Participant	**Age**	**Highest level of education**	**English level**	**YouTube usage**
P1	32	Master of Science	C1	Daily
P2	30	Compulsory school diploma	B1	Multiple times per week
P3*	30	High school diploma	C1	Daily
P4*	34	Master of Arts	C1	Daily
P5	30	Bachelor of Science	C1	Daily
P6*	32	High school diploma	C1	Daily
P7	35	High school diploma	B2	Daily

Table 3.2: An overview of the participants age, highest level of education, english level, and YouTube usage. Entries with an asterisk (*) indicate participants who also participated in the user preference study.

3.4.2 Environment and Setup

The study was conducted in a controlled laboratory environment at the TU Wien [7]. A 32-inch Full HD television connected to a Mac Mini with wireless keyboard and mouse peripherals served as the testing setup. A microphone recorded participant feedback and think-aloud [30] protocols. The researcher sat approximately 2 meters away at a

[7] https://www.tuwien.at/ - last accessed 26.03.2025

Phase	Activity	Duration	Description
1. Consent	Review of consent form	5 min	Participant reviews and signs consent
2. Pre-Assessment	Pre-study survey completion	5 min	Demographic information & YouTube usage
3. Training	Viewing training video	5 min	Task explanation & system introduction
4. Clarification	Question and answers	As needed	Ensure task understanding
5. Main Study	Task execution and observation	45 min	Task sheet provided, Session recording initiated, Researcher note-taking
6. Post-Task Assessment	Questionnaire completion	15 min	Optional break offered, Feedback collection
7. Final Interview	Semi-structured interview	15 min	Audio recording, In-depth feedback collection

Table 3.3: Study protocol to ensure consistency during the user evaluation study.

90-degree angle to the participant (see Appendix: 7 for details). During the user task phase, the keyboard was removed to minimize distractions, as participants only needed the mouse to scroll through comments. The keyboard was reintroduced for the post-task survey.

For the semi-structured interview, the seating arrangement and recording equipment were adjusted to create a conversational atmosphere, fostering a relaxed environment for open and natural responses.

3.4.3 Study Protocol

To ensure methodological consistency, a structured protocol was developed and followed for each session. The procedure was carefully explained to each participant at the beginning of their session, as detailed in Table 3.3.

3.4.4 Pre-Task Survey

The pre-study included a questionnaire designed to collect demographic data and analyze participants' YouTube usage behavior. The survey aimed to understand how participants engage with YouTube, their comment-reading habits, and their attitudes toward automated sentiment analysis tools. By gathering this baseline information, the study ensured a contextualized interpretation of user interactions with the tools. Additionally, the questionnaire explored potential concerns regarding automated comment analysis, providing insights into user expectations and reservations before they interacted with the tools. To assess whether their perspectives evolved after interacting with the tools, the post-task survey included a follow-up question on the perceived usefulness of such a tool. This allowed for a direct comparison between participants' expectations and their actual experiences, providing insights into whether and how their opinions changed. The pre-task survey questions can be found in Appendix: 7.

3.4.5 Training Video

Following the pre-study questionnaire, participants watched a 3-minute 10-second training video [8] created for the study. The video, presented in German with a talking-head overlay and screen recordings of the final tool versions, ensured a consistent introduction without prior tool exposure. Participants could pause or rewind as needed.

The video introduced the interface and workflow: participants were instructed to watch an assigned YouTube video in full, then read the first 20 comments. Sequentially arranged tabs guided them through Tools A, B, and C in a controlled order of increasing complexity. That the complexity would increase was intentionally not mentioned. Tool As explanation focused on its sentiment summary panel and filtering options, including sentiment labels above comments. Tool B introduced the "reasoning" feature, which also, was intentionally, after introduced once, not further commented on (e.g., highlighting benefits) to capture natural reactions of the participants. Tool C highlighted its nuanced classification system beyond basic sentiment categories. After a brief summary, participants were encouraged to verbalize observations during the tasks.

Two participants re-watched the Tool C section for clarity. Comprehension checks revealed no further questions. Participants then received the task sheet, were informed of screen and audio recording, and instructed to use checkboxes to guide their progress, minimizing researcher intervention. The study then proceeded to the user task phase.

3.4.6 User Tasks & Task Structure

Participants followed a structured approach to comment analysis using the task sheet (see Appendix: 7), which consisted of four main tasks, each with three similar sub-tasks:

1. **Initial Content Exposure**: Participants watched a 2-minute YouTube video and read its first 20 comments to establish context.
2. **Systematic Tool Analysis**: Participants re-examined the same comments using each tool (A, B, and C) while verbalizing their thoughts on:
    - Agreement with tool classifications
    - Evaluation of explanations and confidence scores

[8] YouTube Training Video: Training video Tool A, B, C

- Assessment of specialized classification categories and tone analysis

Task duration was not strictly controlled. During the tasks, I took unstructured observational notes, which became more systematic as recurring difficulties and questions emerged over the course of the study user evaluation study. These insights are further explored in Chapter 3.4.

After completing the tasks, participants were offered a break, as they had been engaged for about an hour, including the pre-study questionnaire and training video. This break aimed to prevent fatigue and maintain response quality in the post-task survey and following interview. If declined, the study proceeded directly.

In the main part of the study, in which participants tested the three tools, the tools were presented in a sequential order: first Tool A, then Tool B, finally Tool C. This sequential tool presentation strategy served multiple purposes:

1. Progressive Complexity: Each subsequent tool added new elements while maintaining consistency with at least one aspect of the previous tool, allowing participants to build upon their understanding incrementally.
2. Comparative Analysis: The deliberate ordering facilitated two types of comparisons:
   - Between confidence scores and reasoning (Tool A vs. Tool B)
   - Between traditional and specialized classification schemes (Tools A/B vs. Tool C)
3. Learning Curve Management: By introducing new concepts gradually, the study design helped managing cognitive load while maintaining engagement with increasingly complex analysis methods.

3.4.7 Post-Task Survey

The post-task survey took approximately 15 minutes. Each tool (Tool A, Tool B, and Tool C) was evaluated through 14 questions, totaling 42 questions (see Appendix: 7). The questions followed the testing order—Tool A first, then Tool B, and Tool C—without randomization. Responses were recorded on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree).

One question per tool assessed overall satisfaction, allowing optional comments. The final question in each set asked whether participants views on the usefulness of automated sentiment analysis had changed positively or negatively. The survey consisted of two parts: an overall impression of the tools, followed by individual evaluations based on predefined dimensions.

To ensure clarity, each question included a legend and a screenshot of the corresponding tool, identical to what participants saw during the user tasks with the intention to help them accurately recall the tool they were evaluating.

3.4.8 Thematic Analysis

For the user evaluation study, a thematic analysis was conducted, following the framework by Braun and Clarke [12, 13]. This framework describes thematic analysis as a method for identifying, analyzing, and reporting patterns (themes) within qualitative data. It provides a systematic yet flexible approach, allowing researchers to explore data in depth while maintaining methodological rigor. They propose a reflexive approach, emphasizing the researcher's active role in interpreting the data and viewing subjectivity as a resource rather than a limitation. Key steps include familiarizing oneself with the data, generating initial codes, building categories based on codes, searching for themes based on categories, reviewing themes, defining and naming themes, and producing the final report [12, 13]. This approach ensured coherence between research questions, data collection, and analysis, aligning with the study's goals.

3.4.9 Generating Codes from the Transcript

An Excel sheet generated during transcription listed each interview question alongside participants' responses in paragraph form, forming the basis for coding. If a response mentioned multiple tools (e.g., *"Tool A has... but Tool B is... Tool C is..."*), each sentence was placed in a separate cell to improve coding accuracy.

To the right of each response, the sheet included a coding column and a notes column. Coding was guided by the study hypotheses, allowing paragraphs to receive multiple or no codes. Key phrases exemplifying specific codes were highlighted in blue for easy reference during analysis. The notes column documented thoughts such as *"Future work," "Zitat," (eng. Quote)* or *"Verständnisproblem" (Comprehension Problem).*

A separate "Codelog" sheet tracked all assigned codes after each interview. Codes were labeled descriptively (e.g., "Emojis_not_recognized") but were not yet explicitly defined. Recurring patterns emerged across interviews, though codes were not yet grouped into categories. This process yielded **109** individual codes across all seven interviews.

3.4.10 Consolidating and Refining Codes

All codes were consolidated into a single Excel sheet, where duplicates were removed and codes were grouped and refined, resulting in **39** distinct codes. Using mind-mapping software draw.io [9], these codes were assigned to the corresponding tools based on interviewee comments, providing initial insights into user preferences (see Figure 3.2).

3.4.11 From Codes to Categories

In the next step, codes were logically grouped and clustered to build overarching categories. While some categories, such as "Lack of Explainability" and "Overconfidence in Sentiment Scores", were primarily associated with Tool A, the process resulted in

[9] `https://app.diagrams.net/` - last accessed 23.03.2025

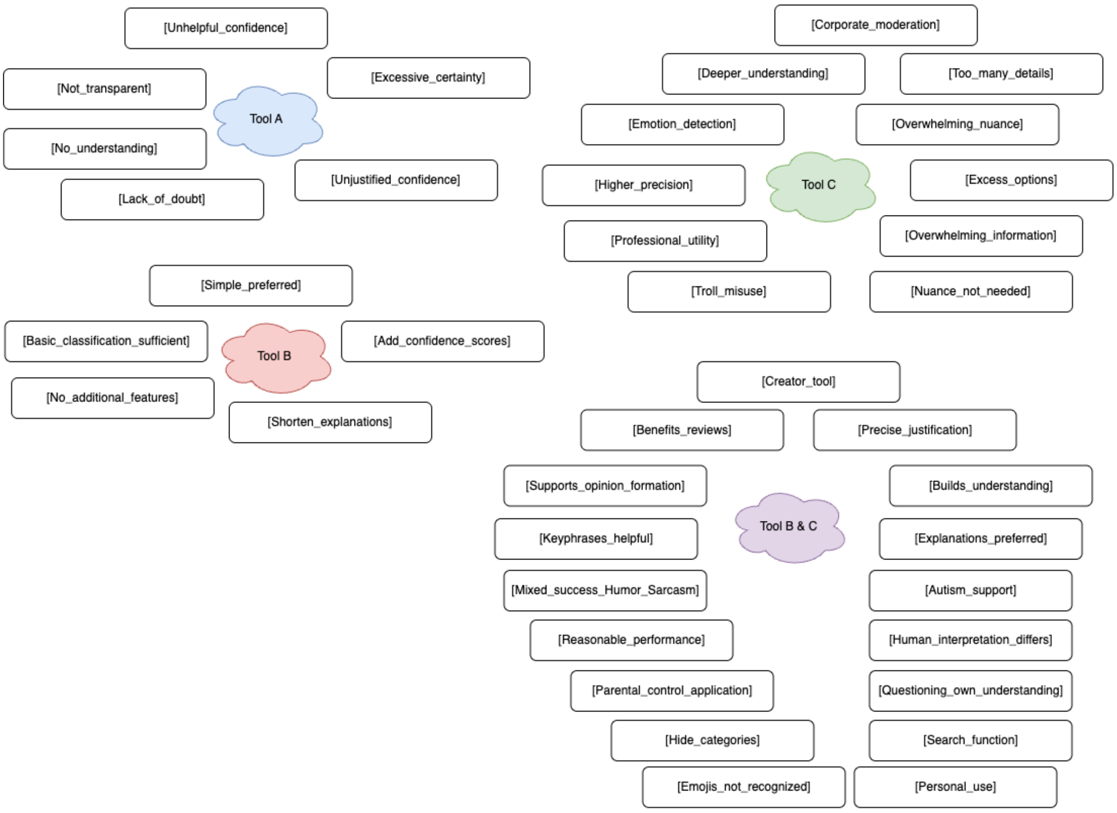

Figure 3.2: The figure shows how the generated codes contextually relate to Tool A, B and C based on what participants commented on in the interviews. It shows that Tool B & C got mostly positively commented on compared to Tool A.

nine overarching categories: *Lack of Explainability, Overconfidence in Sentiment Scores, Simple Categorization, Information Overload, Desired Features, Explanation-Based Approaches, Value of Nuance, Detection Difficulties*, and *Beneficial Users*. Figure 3.3 visually maps codes to their respective categories.

3.4.12 From Categories to Themes

Building on these categories and the research hypotheses, thematic clusters were created, each representing one or more categories. These clusters were reviewed and validated with the thesis advisor to ensure coherence and accuracy. Seven overarching themes emerged: *Confidence & Transparency, Understanding, Target Audience, Information Presentation, Cognitive Load, Feature Requests*, and *Emotion Recognition*. These themes, along with their categories, codes, definitions, and examples, are discussed in detail in Chapter 5. Figure 3.4 visualizes which tool contributed to which category, demonstrating the relationship between tools and thematic development.

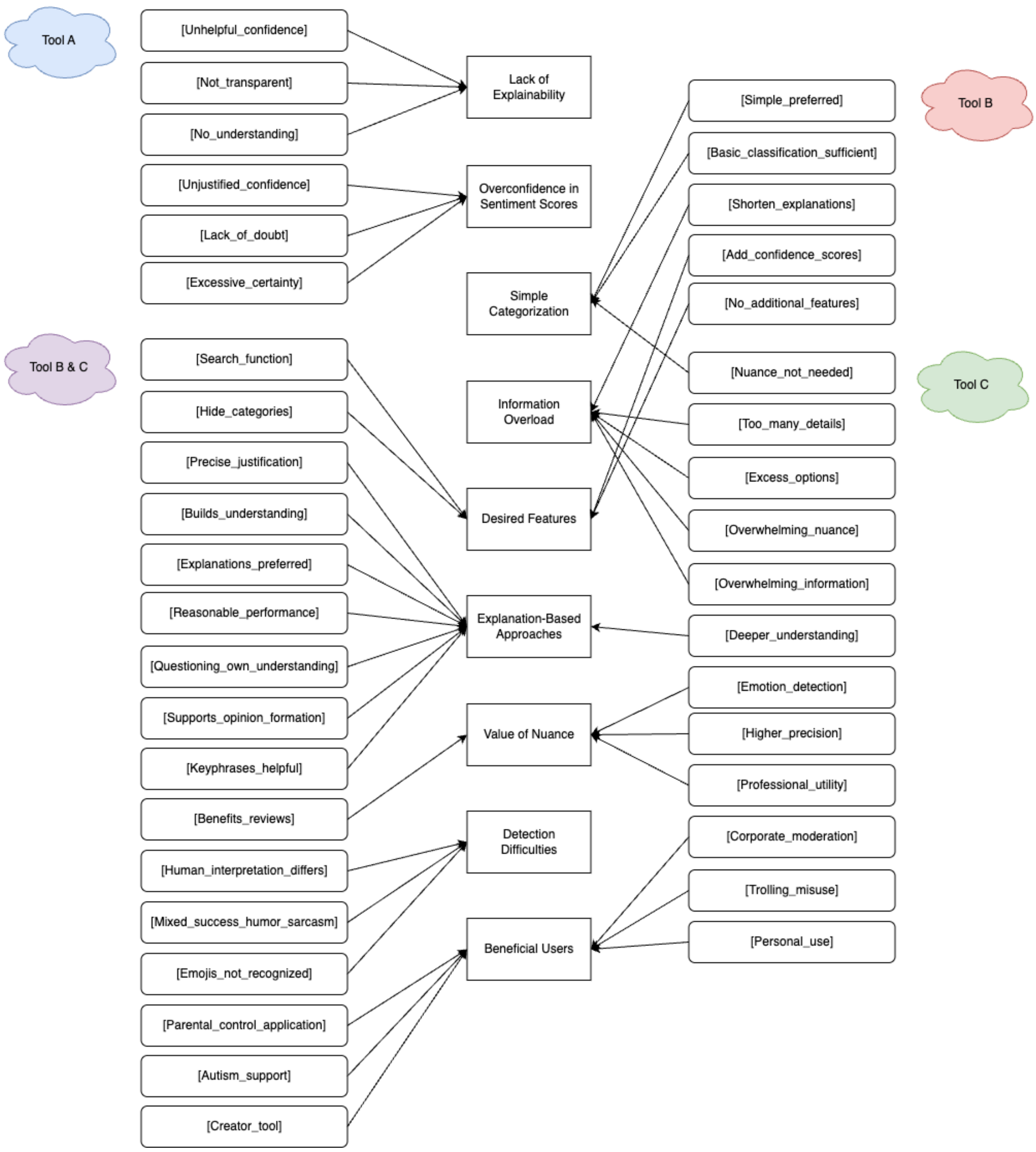

Figure 3.3: The figure shows which codes from which tool cluster contributed to which category. In total, 9 categories (rectangles) were derived from 39 codes (rectangles with rounded corners).

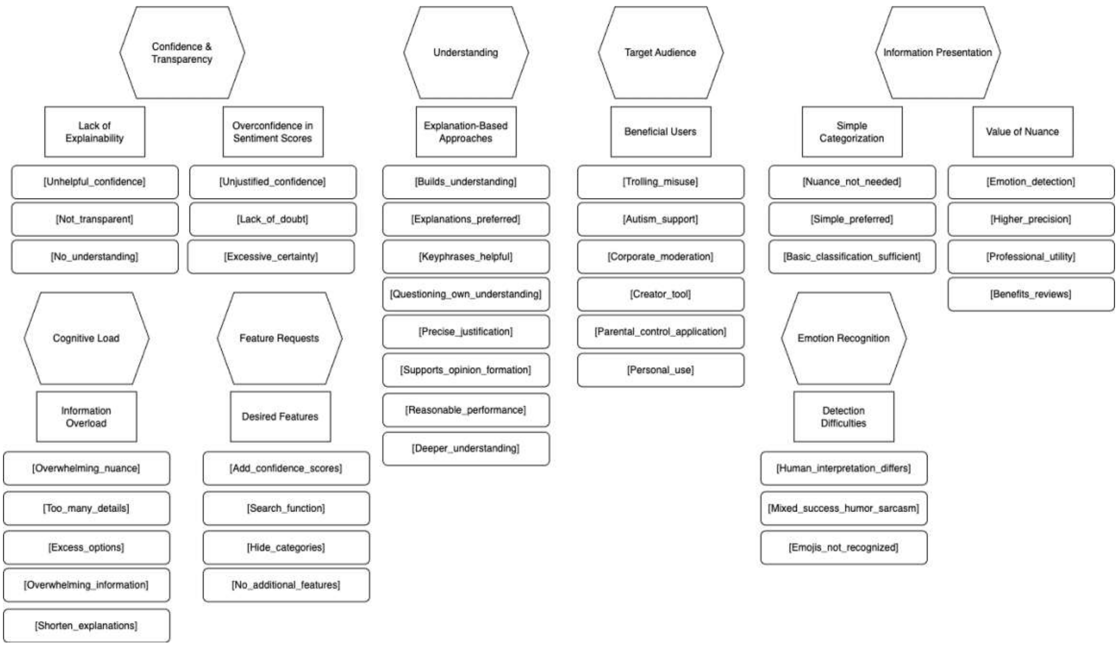

Figure 3.4: The figure shows the overarching themes that got derived from categories. Themes consists of either two (e.g, Confidence & Transparency) or one (e.g., Understanding) underlying category with its assigned codes.

3.4.13 Additional Assessment

To further evaluate model responses, participants completed a third questionnaire on a later day after finishing the study, classifying previously encountered comments (see Section: 4.4) as positive, negative or neutral. This assessment was conducted separately to minimize cognitive load and served as the ground truth for determining which model performed best based on human annotations. Due to LimeSurveys [10] technical limitations in rendering emojis, Google Forms [11] was used instead, as emoji interpretation significantly influences sentiment analysis [100].

Having outlined the studys methodological approach—including its design, user tasks, surveys, and participant recruitment—the next chapter shifts focus to model evaluation. This section examines the quantitative performance of various SLMs and BERT-based models to identify the most suitable model for sentiment classification based on the gathered data.

[10] https://www.limesurvey.org/ - last accessed 23.03.2025
[11] https://docs.google.com/forms/ - last accessed 23.03.2025

CHAPTER 4

Model Evaluation

During the early stages of this thesis in October 2024, it was initially planned to evaluate three SLMs—Llama3.2:1B, Phi-3, and Gemini Nano—by comparing their results with a fine-tuned BERT-based model, trained on the Stanford Sentiment Treebank (*distilbert-base-uncased-finetuned-sst-2-english*).

However, the formal model evaluation was conducted in February 2025. Between October 2024 and February 2025, several new SLMs were released, raising the question of whether the originally proposed models remained valid candidates. For example, Qwen2.5 [1], Phi-4 [2], and DeepSeek-V3 [3] were introduced in December 2024. Additionally, other models, initially overlooked, such as InternLM2.5 [4] and SmolLM2 [5], were later considered for the evaluation as they state to outperform Llama3.2:1b. Just days before the user evaluation study, *mistral-small:22b* [6] was released. As it was the most recent model at the time, it was included in the user evaluation study, replacing the previously planned Phi-4:14b, which itself had already replaced Gemma2:9B—a successor to the originally proposed Gemini Nano. Ultimately, Gemini Nano was entirely excluded from the evaluation due to accessibility constraints, as it requires a dedicated Google Developer Account [7] and participation in a beta testing program [8], which would be impractical for standard users.

In total, 19 SLMs of varying parameter sizes were evaluated, 16 more than initially planned. The model with the highest parameter count was *mistral-small:22b* where

[1] https://github.com/QwenLM/Qwen2.5 - last accessed 23.03.2025
[2] https://techcommunity.microsoft.com/blog/aiplatformblog/introducing-phi-4-microsoft-s-newest-small-language-model-specializing-in-comple/4357090 - last accessed 23.03.2025
[3] https://api-docs.deepseek.com/news/news1226 - last accessed 23.03.2025
[4] https://ollama.com/library/internlm2 - last accessed 23.03.2025
[5] https://ollama.com/library/smollm2 - last accessed 23.03.2025
[6] https://ollama.com/library/mistral-small - last accessed 23.03.2025
[7] https://developers.google.com/ - last accessed 26.03.2025
[8] https://developer.chrome.com/docs/extensions/ai/prompt-api - last accessed 26.03.2025

"22b" means 22 billion parameters, while the smallest was *SmolLM2:135m* where "135M' means 135 million parameters. A characteristic that is fundamental to all evaluated models is that they have to be freely available and can be downloaded via Ollama, a platform that is enabling local inferencing for language models. This is intended to not only ensure reproducibility of the study and the results but to also make it feasible to use the developed extensions "as is" outside of academia.

The primary objective of this evaluation is to determine which of the 19 selected models produces outputs that best align with a pre-established human-annotated ground truth for sentiment classification of 20 YouTube comments. The models' performance was assessed under two conditions: (i) *without context*, also called *zero-context*, where only the comment was provided for classification, and (ii) *with context*, where the classification prompt was augmented with the corresponding YouTube video transcript.

Additionally, the performance of the SLMs were compared to the specialized BERT-based model *distilbert-base-uncased-finetuned-sst-2-english* or in short DistilBERT-sst2, fine-tuned specifically for sentiment analysis. Ultimately, all model classifications were benchmarked against human annotators, who established the ground truth in the sentiment evaluation.

A total of 20 models, including DistilBERT-sst2, were evaluated on the dataset of 20 YouTube comments. The participants of the user evaluation study were responsible for defining the human baseline, as they rated each comment as positive, negative, or neutral during an additional assessment following the user evaluation. These human annotations serve as the reference for evaluating the models' classification accuracy. The next section will describe the chosen models.

4.1 Candidate Models and their Key Attributes

Each model was chosen based on unique strengths that could be helpful for sentiment analysis in a diverse, real-world context such as YouTube comments. The evaluated models are:

Qwen 2.5 - Series (0.5b, 1.5b, 3b, 7b)

- Multilingual support covering over 29 languages, which could be beneficial considering limitations mentioned in other research regarding sentiment analysis and language (see Chapter: 2).
- Enhanced instruction following and the ability to generate structured output (notably in JSON), important for building a reliable processing pipeline based on model responses.
- Model context up to 128K tokens, which, while excessive for sentiment analysis, is a nice-to-have feature.

DeepSeek-R1 (1.5b, 7b, 8b)

- Distilled from the Qwen 2.5 series, retaining many of Qwens core advantages (structured output and multilingual support) in a lighter package.
- Various sizes, with the 1.5b, 7b, and 8b versions being strong candidates for comparison.
- Released in December 2024 which makes it one of the more recent models in this evaluation.

Gemma2 (2b, 9b)

- Lightweight model from Google, emphasizing efficient performance.
- The model card [43] states that the model is well suited for reasoning tasks and that its small size makes it a good candidate for local deployment when limited resources are available.
- Competitors like Llama [9] , Phi-4 [10] and InternLM2.5 [11] challenge the performance of the Gemma2-series.

Llama 3.2 (1b, 3b)

- Optimized for multilingual dialogue and instruction following, with strong performance in summarization and prompt rewriting and capable of running locally on the edge due to its low parameter size.
- Comparable model sizes (1b and 3b) to other competitors on the list and part of the initially proposed evaluation as an open and free to use model.

Phi Model Series (Phi-3.5 3.8b and Phi-4 14b)

- Good for handling long context inputs with robust instruction following and safe output generation.
- Phi-4:14b is positioned as one of the best performing models among smaller LLMs, challenging models with more than twice its size according to benchmarks [12] which sounds promising and could be ideal for capturing subtle sentiment nuances.

InternLM2.5 (1.8b, 7b)

- Noted for outstanding reasoning capabilities according to the model card, particularly in tasks requiring logical or mathematical reasoning.
- Especially the 7b model outperforms Gemma2:9b in 5 out of 6 benchmarks according to their performance evaluation [13].

[9] `https://ollama.com/library/llama3.2` - last accessed 23.03.2025
[10] `https://ollama.com/library/phi4` - last accessed 23.03.2025
[11] `https://ollama.com/internlm/internlm2.5` - last accessed 23.03.2025
[12] `https://ollama.com/library/phi4` - last accessed 23.03.2025
[13] `https://ollama.com/internlm/internlm2.5` - last accessed 23.03.2025

Mistral Small (22b)

- Benchmark-setting performance [14] for small LLMs with robust multilingual support, system prompt adherence, and native function calling (e.g., JSON outputs).
- Specifically designed for use cases like virtual customer service and sentiment analysis.
- Released end of January 2025 and therefore taken into account as the most recent SLM to be evaluated in this study.

SmolLM2 (135m, 360m, 1.7b) [4]

- SmolLM2-1.7B explicitly compares itself on benchmarks with Llama3.2:1b and Qwen2.5:1.5B and reports slightly better scores on a number of benchmarks [15].
- SmolLM2:135m and 360m where considered due to their low parameter size.

Before discussing the evaluation metrics, the following section explains the inclusion criteria that models had to meet to be considered for evaluation in the first place.

4.2 Criteria for Model Selection

In addition to being freely available on the Ollama model hub (to ensure reproducibility), models had to fulfill at least two of the selection criteria in Table 4.1 and needed to be released after 2024 while also being capable to run on reasonable hardware. During this research, a Mac Mini M2 Pro with 32 GB RAM was used. For comparison, at the time of writing, the Mac Mini M4 Pro with 64 GB RAM [16] is publicly available.

4.3 Evaluation Metrics

In evaluating the performance of sentiment classification models, appropriate metrics must be selected to ensure a comprehensive assessment. The *macro-F1 score* was chosen to measure the models performance against human-labeled ground truth. This metric ensures that all classes are treated equally, making it particularly suitable for potentially imbalanced and small datasets over the weighted-F1 score and the micro-F1 score [52].

To assess SLM-specific parameters, the *Performance Metrics Summary* was used, providing insights into the computational efficiency and runtime performance of the locally deployed models.

[14] `https://mistral.ai/news/mistral-small-3` - last accessed 26.03.2025
[15] `https://ollama.com/library/smollm2` - last accessed 23.03.2025
[16] Needs to be selected in the configuration: `https://www.apple.com/shop/buy-mac/mac-mini/apple-m4-pro-chip-with-12-core-cpu-16-core-gpu-24gb-memory-512gb` - last accessed 26.03.2025

Criteria	Description
Multilingual Capability	Essential for processing YouTube's global user comments.
Structured Output (JSON Schema Adherence)	Required for integration with the extension's predefined data structure.
Instruction Following & Context Handling	Important for nuanced sentiment analysis, especially with critical context.
Model Distillation & Efficiency	To evaluate the trade-off between model size and performance. Do more parameter imply better performance?
Domain-Specific Performance	Considered both stated and potentially hidden sentiment analysis capabilities.

Table 4.1: Selection Criteria for Models

4.3.1 Macro-F1 Score

The macro-F1 score is a performance metric used to evaluate classification models, particularly in multi-class or imbalanced datasets. It is calculated as the average of the F1 scores computed independently for each class, ensuring that all classes are treated equally, regardless of their size.

The F1 score itself is the harmonic mean of precision (the proportion of predicted positives that are actually positive) and recall (the proportion of actual positives correctly identified). By balancing precision and recall, the F1 score is especially useful when false positives and false negatives carry similar costs.

By assigning equal weight to each class, the macro-F1 score ensures that minority classes contribute as much to the overall performance measure as majority classes, making it effective in highlighting performance disparities that might otherwise be masked by class imbalances.

4.3.2 Performance Metrics Summary

The *Performance Metrics Summary* as listed in the Ollama documentation[17] provides insights into the following metrics when evaluating models:

1. **Total Duration**:
   - The total time required to complete the task, including model loading, prompt evaluation, and inference.
   - **Example**: `Phi-4` took 13.61 seconds in total.
2. **Load Duration**:

[17] `https://github.com/ollama/ollama/blob/main/docs/api.md` - last accessed 23.03.2025

- The time taken to load the model into memory.
- **Example**: `Phi-4` took 0.01 seconds to load.

3. **Prompt Eval Count**:

   - The number of tokens processed during prompt evaluation (e.g., tokenizing the input).
   - **Example**: `Phi-4` processed 185.55 tokens during prompt evaluation.

4. **Prompt Eval Duration**:

   - The time taken to evaluate the prompt (e.g., tokenizing and preparing the input).
   - **Example**: `Phi-4` took 0.60 seconds for prompt evaluation.

5. **Eval Count**:

   - The number of tokens generated during inference (e.g., the model's output).
   - **Example**: `Phi-4` generated 219.75 tokens during inference.

6. **Eval Duration**:

   - The time taken to generate the output tokens (inference time).
   - **Example**: `Phi-4` took 13.00 seconds for inference.

It has to be noted that *Load Duration* is only relevant when the model is first loaded into memory and, therefore, only affects the time required to analyze the first comment in each model run. To ensure consistency and comparability during model evaluation, all models were preloaded in Ollama before analysis (see Listing: 4.1). Additionally, to guarantee result reproducibility, all models were configured with the same *option flags* during inference: (i) the *seed* parameter was set to 1 (ii) the *temperature* was set to 0 (iii) The maximum number of processable tokens (*num_ctx*) was set to 8192 tokens, which corresponds to approximately 6,000 English words [18] or 13 pages [19] of standard font style (Arial) and font size (12px) text.

Listing 4.1: Request to keep the model preloaded in memory and set option flags for result reproducibility

```
payload = {
    'model': model,                  # e.g: Phi-4
    'keep_alive': -1,                # keep model in memory
    'stream': False,
    'options': {
        'temperature': temperature,  # set to 0
```

[18] https://platform.openai.com/tokenizer - last accessed 23.03.2025
[19] https://wordcounter.net/words-per-page - last accessed 23.03.2025

```
        'seed': seed,                   # set to 1
        'num_ctx': num_ctx              # set to 8192
    }
}
```

4.4 The Dataset

To evaluate the models, a dataset of 20 comments was collected from a YouTube video [20]. The selected comments represent the top 20 comments of the YouTube video at the time of scraping (11.02.2025) and capture a diverse range of linguistic features commonly found in online discussions, including capitalized words, abbreviations (e.g. “GOAT” = Greatest Of All Time or “CGI” = Computer-Generated Imagery) heavy use of exclamation marks (e.g “!!!”) and direct quotes from the video (e.g “I am so good” or “Shuffles legit obviously”).

Notably, comments 4, 5, 7, 11, 15, and 18 incorporate emojis to emphasize their statements. The dataset also includes humorous and sarcastic remarks, adding an additional layer of complexity for sentiment analysis. Table 4.2 provides a complete overview of the selected comments.

To assess the accuracy of sentiment classification, this dataset required reliable human annotation, which is discussed in the next section.

[20] https://www.youtube.com/watch?v=rnCjM_sovp0 - last accessed 23.03.2025

No.	Comment
1	Leaving. Very impressive but getting old. Peace out.
2	Hey Jason, love you and your videos. My name is XXX and I'm in the US army and currently deployed in Iraq (me in pfp). Your videos have gotten me through a lot of tough days of deployment so first off thank you for entertaining me. But I have a request, I'm scheduled to come home on June 18th so if you could do a trick, any trick you like with the 6, 1, and 8 it would make my whole week. I wish you nothing but more success and growth on your channel and please keep posting so I don't die of boredom out here. I plan on going to a show when I get back this summer so post those dates as soon as you know them! All love from Iraq!
3	He even trash talks himself truly no one is safe
4	Welcome to the hustle, indeed! 😅 I love how the "includes paid promotion" tag was there. I don't even know what's real anymore. But you are still the goal.
5	"I am so good" 😂😂😂
6	It doesn't matter how many times I watch these videos or how many times I think I might've seen something... This guy right here is hands down the best I've ever seen. It's not even close. And he seems like he'd be a helluva guy to hang out with! Thanks for all the awesome content!
7	Nicely done and she paid 500 for an advertisement. 😂
8	Saw Jason live in Boca Raton. He is even better live. Definitely the GOAT!!
9	Always love the waterfall trick! Truly impressive
10	I have been watching you very closely do these tricks and I can never spot how in the world you pull these tricks off. So after careful analysis, I have decided that Magic is 100% real and you my friend have mastered it!
11	This man's CGI machine must be burning 😂
12	Another great vid. His CGI crew must be up to half a dozen employees by now to be able to keep up with all his amazing tricks Genius is expensive
13	Hardest thing for Mr Jason - talking to the wall like there is actually someone standing
14	Respect for leaving in the blooper
15	Leap out of the deck??!! 😂 cgi machines ready 😂😂😂 Jason you leave me stunned my friend!!! you are so MASTER GOAT!!! ❤️❤️ Greetz Spiecull. 🎉
16	I love when someone who is the best at something is arrogant. You've certainly earned the right.
17	Waterfall with two cards. Absolutely incredible!
18	I really enjoy this channel. Your card manipulation skills are mind blowing. Take care eh ✌️🇨🇦
19	Aha!mm i figuree it out ...If you look at the dark silver car thats just behind Jasons head on the book shelf payclose attention to the angle of the green case .. then after he touches his face roughly around 42 seconds in you will notice that i have no idea how he does it
20	"Shuffles legit obviously."

Table 4.2: YouTube Comments Collection

4.5 The Role of Human Annotators

To ensure a reliable ground truth for sentiment classification, the dataset referenced in Section 4.4 was annotated by participants in a controlled experimental setting (as detailed in Chapter: 5). These annotators, who actively engaged with the associated YouTube video, provided insights by incorporating their understanding of context, sarcasm, and humor—elements that are often challenging for models to interpret.

Since sentiment is not always strictly positive or negative, a neutral category was introduced, resulting in a three-way classification system (positive, negative, neutral). This human-labeled dataset serves as a baseline for evaluating model performance and identifying potential gaps in automated sentiment analysis.

To examine how well sentiment models align with human perception, the next section compares this dataset against DistilBERT-sst2, a BERT-based model fine-tuned for sentiment analysis.

4.6 Comparison with a BERT-based Model Fine-Tuned for Sentiment Analysis

In this study, *distilbert-base-uncased-finetuned-sst-2-english* is used as the BERT-based model for sentiment analysis, as it offers a high accuracy rate (91.3%) for its size compared to its bigger base model, *BERT bert-base-uncased* (92.7%) [21].

This BERT-based model is a fine-tuned checkpoint of *DistilBERT-base-uncased*[97], trained on the Stanford Sentiment Treebank dataset[22], fine-tuned for binary (positive/negative) sentiment analysis and provides a confidence score for its results. However, several limitations [23] exist:

- The fine-tuning dataset does not include emojis and is limited to English text.
- The model is trained to classify short text up to a sequence length of 128 tokens, the base model is capable of processing 512 tokens.
- The model may exhibit biases in sentiment classification. The authors note that it could 'produce biased predictions that target underrepresented populations.' For example, in the sentence "This film was filmed in COUNTRY" the model assigns radically different probabilities for the positive label depending on the country—0.89 for France but only 0.08 for Afghanistan—despite no explicit semantic difference [54].

The next section evaluates how well the models perform in classifying sentiment and how they compare to human annotations.

4.7 Binary Classification Evaluation

To effectively assess sentiment classification performance, the dataset and models must be adapted for binary classification. This section outlines the necessary adjustments to the dataset, the preparation of the SLMs for binary classification, and the resulting performance. The findings highlight trends in model accuracy and the impact of contextual information on classification outcomes.

4.7.1 Adjusting the Dataset for Binary Classification

To enable a direct comparison between model predictions and human annotations, the dataset (see Section: 4.4) was adjusted from a three-way classification (positive, negative, neutral) to a binary classification (positive/negative). This modification ensures consistency when evaluating model performance against human-labeled data.

[21] `https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english` - last accessed 23.03.2025

[22] `https://huggingface.co/datasets/stanfordnlp/sst2` - last accessed 23.03.2025

[23] `https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english` - last accessed 23.03.2025

No.	Comment	Original Human Annotation	Modified Human Annotation
1	Leaving. Very impressive but getting old. Peace out.	NEUTRAL	NEGATIVE*
13	Hardest thing for Mr Jason - talking to the wall like there is actually someone standing	NEUTRAL	POSITIVE*
20	"Shuffles legit obviously."	NEUTRAL	POSITIVE*

Table 4.3: This table shows which comments had to be modified for proper binary evaluation. From 20 comments, 3 are affected and are marked with an asterisk (*).

A review of the human-annotated dataset revealed three comments requiring reclassification (see Table: 4.3). In two cases (No. 13 and 20), votes were split between positive and neutral, with no negative classifications. Since annotators who selected neutral may have leaned toward positive in a binary setting, these comments were reassigned as positive. Similarly, one comment (No. 1) had a split between neutral and negative, with no positive votes, leading to its reassignment as negative.

For clarity, modified annotations are marked with an asterisk (e.g., Positive*). This adjustment allows for a more accurate binary evaluation while preserving the original sentiment distribution as closely as possible. The next section details how the SLMs were prepared for binary classification.

With the dataset adjusted for binary classification, the next step involved preparing the SLMs to perform sentiment analysis. The following section outlines the prompting strategies used to instruct the models, including both *zero-context* classification and an enhanced approach that incorporated video transcript context.

4.7.2 Preparing the SLM for Binary Classification

In the first step, the SLMs were instructed to perform a binary classification of the comments using the prompt shown in Listing 4.2 for the previously extracted comment (*{comment}*) from the YouTube comment section.

Listing 4.2: Prompt for binary zero-context classification.

```
prompt = f"""
Task: Analyze the sentiment of the YouTube comment and provide
   detailed reasoning.

Instructions:
    1. Analyze the comment's emotional tone, word choice, and
       context
    2. Classify the sentiment as POSITIVE or NEGATIVE
    3. Provide clear reasoning that includes:
        - Tone analysis
        - Key phrases or words that influenced your decision
        - Context consideration from the transcript if available
    4. Format your response as a JSON object with two fields:
        - sentiment: Your classification (POSITIVE/NEGATIVE)
        - reasoning: Your detailed analysis
```

```
YouTube Comment to analyze:
"{comment}"
"""
```

In an additional evaluation, the SLMs were provided with the transcript (see Appendix 7) of the YouTube video as context (*{context}*) to assess whether model performance improves with the inclusion of contextual data. The adjusted prompt is shown in Listing 4.3.

Listing 4.3: Prompt for binary classification with context.

```
prompt = f"""
Task: Analyze the sentiment of the YouTube comment and provide
   detailed reasoning.
Video Transcript Context: {context}

Instructions:
    1. Analyze the comment's emotional tone, word choice, and
       context
    2. If transcript is provided, consider the video context in
       your analysis
    3. Classify the sentiment as POSITIVE or NEGATIVE
    4. Provide clear reasoning that includes:
        - Tone analysis
        - Key phrases or words that influenced your decision
        - Context consideration from the transcript if available
    5. Format your response as a JSON object with two fields:
        - sentiment: Your classification (POSITIVE/NEGATIVE)
        - reasoning: Your detailed analysis

YouTube Comment to analyze:
"{comment}"
"""
```

Each of the SLMs was also instructed to follow a specified output schema (see Listing: 4.4) to ensure that the results were compatible with the frontend components.

Listing 4.4: Binary classification response schema.

```
default_schema = {
    "type": "object",
    "properties": {
        "sentiment": {"enum": ["POSITIVE", "NEGATIVE"]},
        "reasoning": {"type": "string"}
    },
    "required": ["sentiment", "reasoning"]
}
```

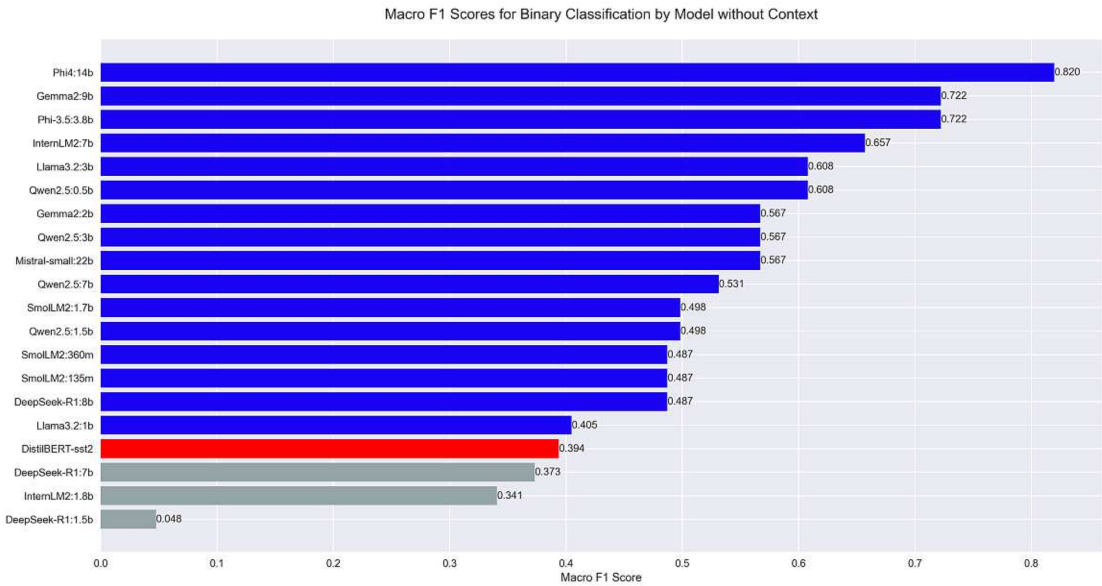

Figure 4.1: **Macro-F1 Scores of Language Models in Binary Zero-Context Evaluation.** This figure compares the performance of 20 language models on binary sentiment classification tasks (positive/negative) without additional context (e.g., no transcript) on 20 YouTube comments. The y-axis represents the language models, while the x-axis shows the macro-F1 scores. Phi-4:14b achieves the highest score (0.820), followed by Phi-3.5:3.8b and Gemma2:9b. In contrast, models like DeepSeek-R1:1.5b and InternLM2.5:1.8b show significantly lower performance. Blue bars highlight models which perform better than DistilBERT-sst2, which is highlighted in red.

4.7.3 Findings: Binary Classification

After the human ground truth of the dataset was adjusted and processed with all the models, it was observed which model most closely represented the human ground truth. When model performance was evaluated without the YouTube transcript as additional context, DistilBERT-sst2 was found to rank among the lowest based on the macro-F1 score (see Figure: 4.1). In contrast, SLMs such as Phi-4:14b, Phi-3.5:3.8b, and Gemma2:9b were observed to perform best, showing closer alignment with the human baseline.

When comparing the human-annotated results with those of DistilBERT-sst2, it is observed that the BERT-based model classifies six comments (No. 3, 11, 12, 13, 19, 20) as negative, while the human annotators classified only one comment as *NEGATIVE** (No. 1) (see Table: 4.4).

Further, examining the sentiment distribution (see Figure: 4.2) of models such as InternLM2.5:1.8b, DeepSeek-R1-7b, and Llama3.2:1b, the data shows that they classify more than half of the comments as negative. DeepSeek-R1:1.5b rates all comments as negative, while SmolLM2:135m, SmolLM2:360m, and DeepSeek-R1:8b classify every comment as positive.

No.	Comment	Human Annotator	DistilBERT-sst2
1	Leaving. Very impressive but getting old. Peace out.	NEGATIVE*	POSITIVE
2	Hey Jason, love you and your videos. My name is XXX and I'm in the US army and currently deployed in Iraq (me in pfp). Your videos have gotten me through a lot of tough days of deployment so first off thank you for entertaining me. But I have a request, I'm scheduled to come home on June 18th so if you could do a trick, any trick you like with the 6, 1, and 8 it would make my whole week. I wish you nothing but more success and growth on your channel and please keep posting so I don't die of boredom out here. I plan on going to a show when I get back this summer so post those dates as soon as you know them! All love from Iraq!	POSITIVE	POSITIVE
3	He even trash talks himself truly no one is safe	POSITIVE	NEGATIVE
4	Welcome to the hustle, indeed! 😅 I love how the "includes paid promotion" tag was there. I don't even know what's real anymore. But you are still the goal.	POSITIVE	POSITIVE
5	"I am so good" 😂😂😂	POSITIVE	POSITIVE
6	It doesn't matter how many times I watch these videos or how many times I think I might've seen something... This guy right here is hands down the best I've ever seen. It's not even close. And he seems like he'd be a helluva guy to hang out with! Thanks for all the awesome content!	POSITIVE	POSITIVE
7	Nicely done and she paid 500 for an advertisement. 😂	POSITIVE	POSITIVE
8	Saw Jason live in Boca Raton. He is even better live. Definitely the GOAT!!	POSITIVE	POSITIVE
9	Always love the waterfall trick! Truly impressive	POSITIVE	POSITIVE
10	I have been watching you very closely do these tricks and I can never spot how in the world you pull these tricks off. So after careful analysis, I have decided that Magic is 100% real and you my friend have mastered it!	POSITIVE	POSITIVE
11	This man's CGI machine must be burning 😂	POSITIVE	NEGATIVE
12	Another great vid. His CGI crew must be up to half a dozen employees by now to be able to keep up with all his amazing tricks Genius is expensive	POSITIVE	NEGATIVE
13	Hardest thing for Mr Jason - talking to the wall like there is actually someone standing	POSITIVE*	NEGATIVE
14	Respect for leaving in the blooper	POSITIVE	POSITIVE
15	Leap out of the deck??!! 😂 cgi machines ready 😂😂😂 Jason you leave me stunned my friend!!! you are so MASTER GOAT!!! ❤️❤️ Greetz Spiecull. 🎉	POSITIVE	POSITIVE
16	I love when someone who is the best at something is arrogant. You've certainly earned the right.	POSITIVE	POSITIVE
17	Waterfall with two cards. Absolutely incredible!	POSITIVE	POSITIVE
18	I really enjoy this channel. Your card manipulation skills are mind blowing. Take care eh ✌️🇨🇦	POSITIVE	POSITIVE
19	Aha!mm i figuree it out ...If you look at the dark silver car thats just behind Jasons head on the book shelf payclose attention to the angle of the green case .. then after he touches his face roughly around 42 seconds in you will notice that i have no idea how he does it	POSITIVE	NEGATIVE
20	"Shuffles legit obviously."	POSITIVE*	NEGATIVE

Table 4.4: This table shows all YouTube comments with their sentiment ratings from the human annotators (modified) and from DistilBERT-sst2 next to each other. Discrepancies can be seen in: No. 3, 11, 12, 13, 19, 20

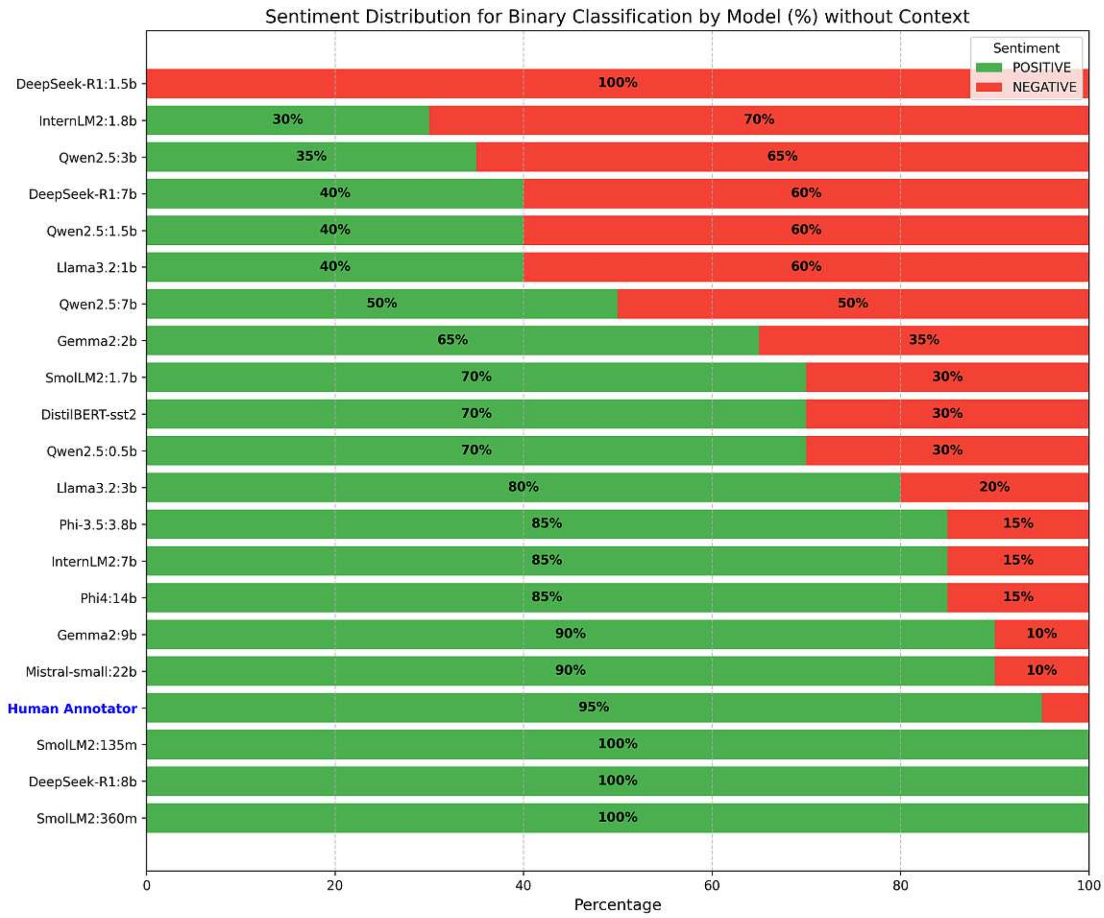

Figure 4.2: **Binary Sentiment Distribution by Model (%) in Zero-Context Evaluation.** Comparing 20 language models' performance on sentiment classification tendencies of 20 YouTube comments with no additional transcript with human annotators. The y-axis shows the models, the x-axis shows how many of the comments were classified as either positive or negative in percent. The chart highlights variations in sentiment classification, with some models exhibiting a strong biases toward positive sentiment (SmolLM2:135m, SmolLM2:360m, DeepSeek-R1:8b) while others have a strong bias towards negativity (DeepSeek-R1:1.5b, DeepSeek-R1:7b, InternLM2.5:1.8b, Llama3.2:1b) and others provide a sentiment analysis closer to human annotations (InternLM2.5:7b, Gemma2:9b, Phi-3.5:3.8b & Phi-4:14b).

In the binary classification *zero-context* evaluation, Phi-4:14b performed the best, achieving a macro-F1 score of 0.82 and a sentiment distribution of 18 positive and 2 negative comments, compared to the human ground truth (19 positive and 1 negative). DistilBERT-sst2 ranked 17th out of 20, with a macro-F1 score of 0.394 and a sentiment distribution of 14 positive and 6 negative comments. It should be noted that, due to the small dataset, any model that classified all comments as positive (e.g., SmolLM2:135m) outperformed DistilBERT-sst2 in binary classification.

Building on the binary classification results from the previous evaluation, the next step was to explore the effect of providing additional context, specifically the video transcript, on model performance. This context-enhanced binary classification evaluation aimed to assess whether the inclusion of contextual information could improve the classification accuracy (macro-F1 score).

4.7.4 Context-Enhanced Binary Classification

The second evaluation maintained the binary classification paradigm but incorporated additional contextual information in the form of the video transcript. The goal was to measure the impact of providing contextual information on the macro-F1 score, with the assumption that classification accuracy would increase. The evaluation was conducted exclusively with the SLM-based models and compared to the *zero-context* results from DistilBERT-sst2. The BERT-based model was not used for contextual analysis, as it cannot process token sequences longer than 128 tokens [24]. For reference, the transcript alone consists of 366 tokens when tokenized with the *AutoTokenizer* for *distilbert-base-uncased-finetuned-sst-2-english* [25]. Even the base models maximum sequence length of 512 tokens was insufficient when combining the transcript (366 tokens) with *comment No. 2* (156 tokens), exceeding the limit at 524 tokens.

Enabling token truncation to fit within the models constraints resulted in the loss of the final portion of the comment, which could contain important sentiment information. Additionally, when the transcript was pre-pended or appended to the comment, the classification consistently produced a negative sentiment. Due to these limitations, the model was ultimately used to analyze the comment without additional context through out the model evaluation.

Regarding the SLMs: augmenting the classification with contextual information yielded mixed results. Eight models showed a decrease in performance (Qwen2.5:0.5, 1.5b, 3b, 7b, Phi-4:14b, Llama3.2:1b, InternLM2.5:1.8b, Gemma2:2b), six models remained consistent in their classification (SmolLM2:135m, 360m, Phi-3.5:3.8b, DeepSeek-R1:1.5b, 8b, 7b), while others benefited from the additional contextual cues (SmolLM:1.7b, mistral-small:22b, Llama3.2:3b, InternLM2.5:7b, Gemma2:9b). DistilBERT-sst2 ranked 13th out

[24] `https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english` - last accessed 23.03.2025

[25] `https://huggingface.co/docs/transformers/main/model_doc/auto#transformers.AutoTokenizer` - last accessed 23.03.2025

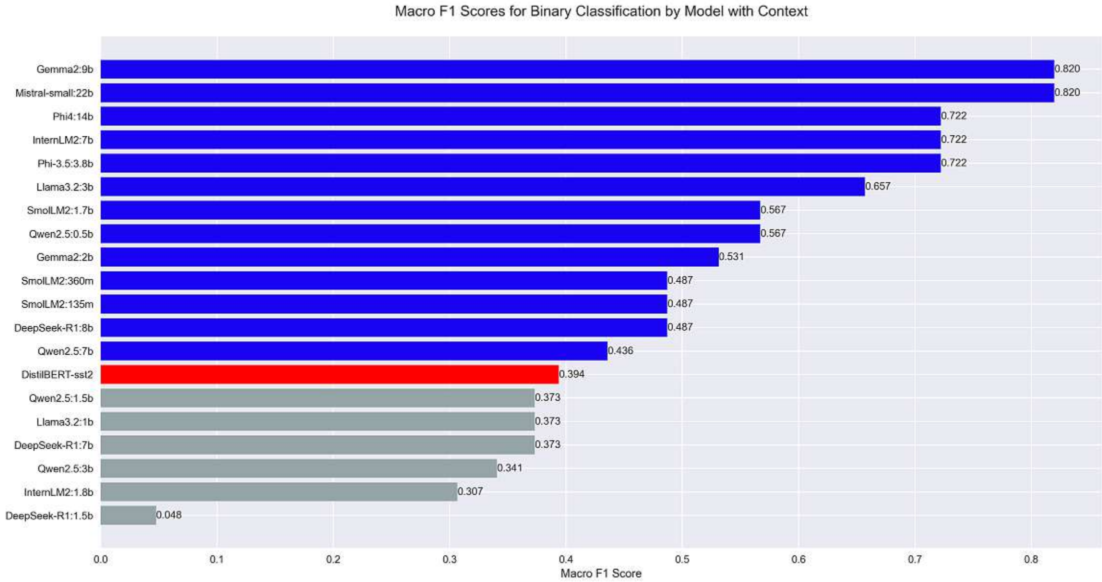

Figure 4.3: **Macro-F1 Scores of Language Models in Binary Classification with Context.** Comparison of 20 language models' performance on binary sentiment classification tasks (positive/negative) using contextual information, the transcript, on 20 YouTube comments. The y-axis shows the language models, and the x-axis represents the macro-F1 score. Mistral-small:22b and Gemma2:9b achieve the highest scores (0.820), followed by Phi-4:14b, Phi-3.5:3.8b, and InternLM2.5:7b (0.722). Other models, like DeepSeek-R1:1.5b and InternLM2.5:1.8b, exhibit significantly lower performance. Blue bars highlight models which perform better than DistilBERT-sst2, which is highlighted in red.

of 20, with models such as Qwen2.5:3b, 1.5b, and Llama3.2:1b performing worse *with context* than without (see Figure: 4.3).

The most significant improvement was observed in the model mistral-small:22b, which increased its macro-F1 score from 0.567 to 0.82, suggesting that the addition of context positively impacted its performance.

When examining the sentiment distribution for binary classification *with context* (Figure 4.4), it is evident that models like SmolLM2:135m, SmolLM2:360m, and DeepSeek-R1:8b continue to classify every comment as positive, while DeepSeek-R1:1.5b rates all comments as negative. Compared to the previous sentiment distribution, it appears these models did not benefit from the additional context.

The top-performing SLMs in this task, Gemma2:9b and mistral-small:22b, correctly classified the first comment as negative, as did the human annotator. However, they also classified comment No. 13, *"Hardest thing for Mr Jason - talking to the wall like there is actually someone standing"* as negative. The reasoning behind that can be seen in Listing 4.5 for Gemma2:9b and in Listing 4.6 for mistral-small:22b.

Listing 4.5: Response from Gemma2:9b for classifying comment No. 13

```
The comment expresses a negative sentiment towards Mr. Jason's
   communication style.

- Tone analysis: The phrase 'Hardest thing for Mr. Jason' implies
   difficulty and struggle, suggesting a negative connotation. The
    comparison to 'talking to the wall' conveys a sense of
   ineffectiveness and lack of engagement from the other party,
   further reinforcing the negative sentiment.

- Key phrases/words: 'Hardest thing,' 'talking to the wall,'
   convey a sense of frustration and futility.

- Context consideration: While we don't have the full video
   context, the comment suggests Mr. Jason is speaking without
   receiving a proper response or engaging in meaningful
   conversation.
```

Listing 4.6: Response from mistral-small:22b for classifying comment No. 13

```
The emotional tone of the comment is critical and mocking. The
   word choice, particularly 'talking to the wall,' suggests that
   the speaker finds Mr. Jason's behavior ridiculous or
   ineffective. This phrase implies that Mr. Jason is speaking as
   if there is an audience when there isn't one, which could be
   seen as a form of delusion or lack of self-awareness. The
   context from the transcript shows Mr. Jason performing a card
   trick and mentioning a dentist, but the comment focuses on his
   mannerisms rather than the content of the video. Overall, the
   comment is meant to ridicule Mr. Jason's behavior.
```

Notably, comment No. 13 was rated from all models as negative, except the ones, which rated all comments as positive. In general, the data suggests that providing additional context pushes some of the models more towards a negative classification although it was expected to have the opposite effect.

4.7.5 Binary Classification Discussion

In conclusion, 80% of the evaluated SLMs outperformed DistilBERT-sst2 in *zero-context* binary sentiment classification and 65% in *with context*. More SLMs performed better without additional context, whereas models with higher macro-F1 scores generally performed better when provided *with context*. The best performing SLMs *with context* were Gemma2:9b and mistral-small:22b, while Phi-4:14b led in *zero-context* performance, all achieving a macro-F1 score of 0.82.

From a researcher's perspective, Gemma2:9b would be considered the “winner” in this setting. It demonstrated better instruction-following, particularly in reasoning based

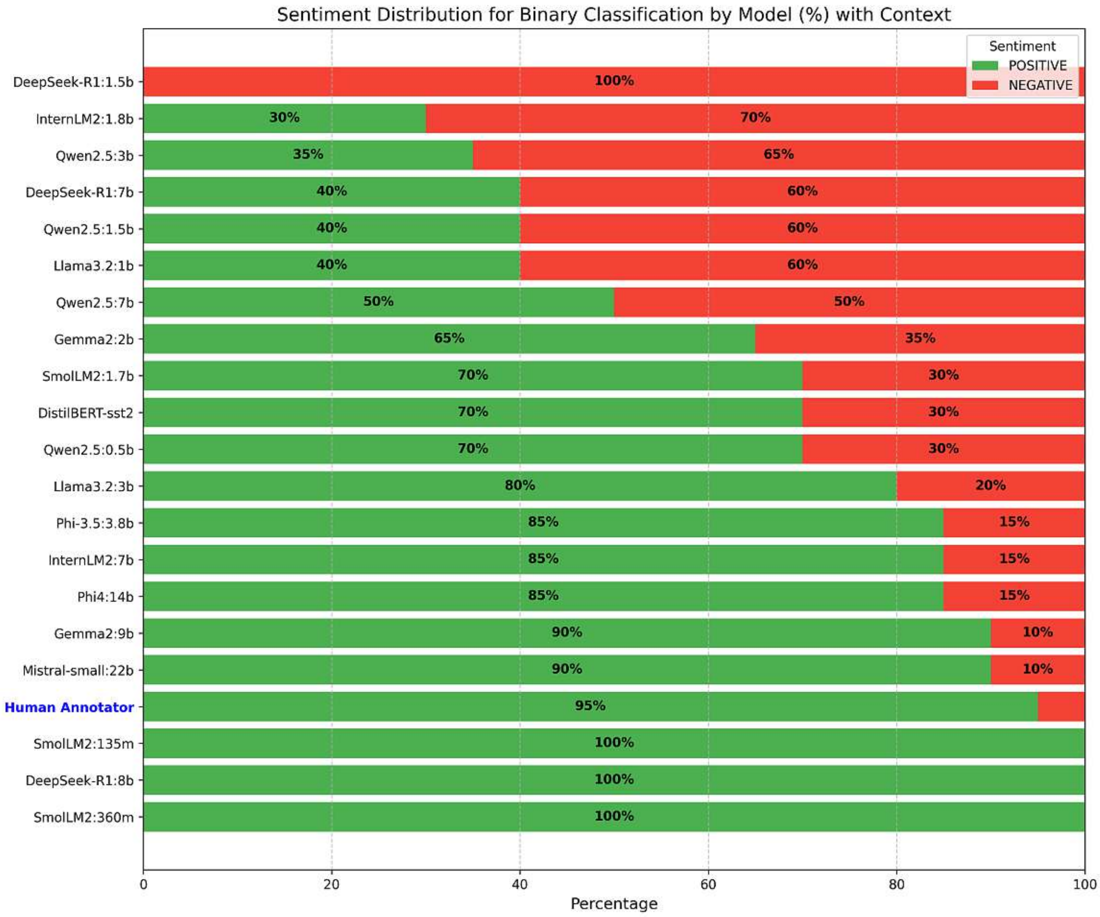

Figure 4.4: **Binary Sentiment Distribution by Model (%) with Context Evaluation.** Comparing 20 language models' performance on sentiment classification tendencies of 20 YouTube comments including a transcript with human annotators. The y-axis shows the models, the x-axis shows how many of the comments were classified as either positive or negative in percent. The chart highlights variations in sentiment classification, with some models exhibiting a strong biases toward positive sentiment (SmolLM2:135m, SmolLM2:360m, DeepSeek-R1:8b) while others have a strong bias towards negativity (DeepSeek-R1:1.5b, Llama3.2:1b, InternLM2.5.1:8b) and others provide a sentiment analysis closer to human annotations (mistral-small:22b, Gemma2:9b, InternLM2.5:7b, Phi-4:14b & Phi-3.5:3.8b).
Note: There was no context (transcript) provided for DistilBERT-sst2.

on the prompt. Furthermore, with only 5.4 GB in disk space (quantized to Q4_0), Gemma2:9b [26] is significantly more efficient compared to the larger mistral-small:22b [27] (quantized to Q4_0), which requires 13 GB of disk space.

As shown in the previous section, it was observed that SLMs outperformed DistilBERT-sst2 in binary classification. The classification is now expanded to include a *Neutral* category, as many comments were rated as neutral by human annotators. In the following section, SLMs are evaluated in a three-way classification system (positive/neutral/negative) *without context*.

4.8 Three-way Classification without Context

As was observed, SLMs were found to outperform DistilBERT-sst2 in binary classification. Participants mentioned during the interviews of the user preference study that comments are not only positive or negative but could also be just *neutral*. Therefore the *NEUTRAL* category was introduced during the user evaluation study and for the human annotators. Although some comments received mixed classifications, the mode (the most frequently assigned label) was used to determine the final sentiment rating. Consequently, the majority of comments were classified as either positive or neutral.

To evaluate whether better alignment with the human ground truth could be achieved, the classification schema was expanded to include the neutral category as well, resulting in a three-way classification system (positive/neutral/negative). The SLMs were first evaluated without contextual information. The prompt and response schema were adjusted to incorporate the *NEUTRAL* class, as shown in Listings 4.7 and 4.8.

Listing 4.7: Prompt for three-way zero-context classification.

```
prompt = f"""
Task: Analyze the sentiment of the YouTube comment and provide
    detailed reasoning.

Instructions:
    1. Analyze the comment's emotional tone, word choice, and
       context
    2. Classify the sentiment as POSITIVE, NEUTRAL, or NEGATIVE
    3. Provide clear reasoning that includes:
       - Tone analysis
       - Key phrases or words that influenced your decision
       - Context consideration from the transcript if available
    4. Format your response as a JSON object with two fields:
       - sentiment: Your classification (POSITIVE/NEUTRAL/
         NEGATIVE)
       - reasoning: Your detailed analysis
```

[26] https://ollama.com/library/gemma2 - last accessed 26.03.2025
[27] https://ollama.com/library/mistral-small:22b - last accessed 26.03.2025

```
YouTube Comment to analyze:
"{comment}"
"""
```

Listing 4.8: Three-way classification response schema.

```
default_schema = {
    "type": "object",
    "properties": {
        "sentiment": {"enum": ["POSITIVE", "NEUTRAL", "NEGATIVE"
           ]},
        "reasoning": {"type": "string"}
    },
    "required": ["sentiment", "reasoning"]
}
```

Even though DistilBERT-sst2 is unable to classify comments as *Neutral*, it will still be included in the figures for comparison.

Looking at the first chart displaying the macro-F1 score (see Figure: 4.5), it can be seen that DistilBERT-sst2 shares the same score as Gemma2:2b, both scoring 0.280, thus ranking 14th/13th out of the 20 models. In general, the macro-F1 scores of the models are lower compared to the binary classification evaluation, with Phi-3.5:3.8b ranking the highest at 0.490, compared to the previous highest score of 0.820. As shown in Figure 4.5, the overall performance of the models, as measured by the macro-F1 score, decreased compared to the results from the binary classification task.

The highest-performing models in this evaluation are Phi-3.5:3.8b (0.490) and Qwen2.5.3b (0.429). Most other models score between 0.2 and 0.4, with a few notable underperformers, including DeepSeek-R1:1.5b (0.087), Llama3.2:1b (0.152), and InternLM2.5:1.8b (0.152).

When examining the sentiment distribution in Figure 4.2, it is apparent that human annotators did not classify any comment as negative. While some models assigned neutral classifications, the data indicates a tendency toward negative classifications rather than neutral. The best-performing models in this context were Phi-3.5:3.8b and Phi-4:14b. In contrast, SmolLM2:135m and 360m, as well as DeepSeek-R1:1.8b, classified all comments as positive, while DeepSeek-R1:1.5b categorized all comments as neutral.

Overall, Figure 4.6 highlights the variability in sentiment classification strategies across models, with some exhibiting strong biases toward positive sentiment, while others demonstrate a more balanced or nuanced distribution.

Having explored the performance of models in the three-way classification *without context*, the next step is to incorporate additional contextual information through the video transcript. This evaluation aims to determine whether the inclusion of context improves classification accuracy and aligns more closely with human-like judgments.

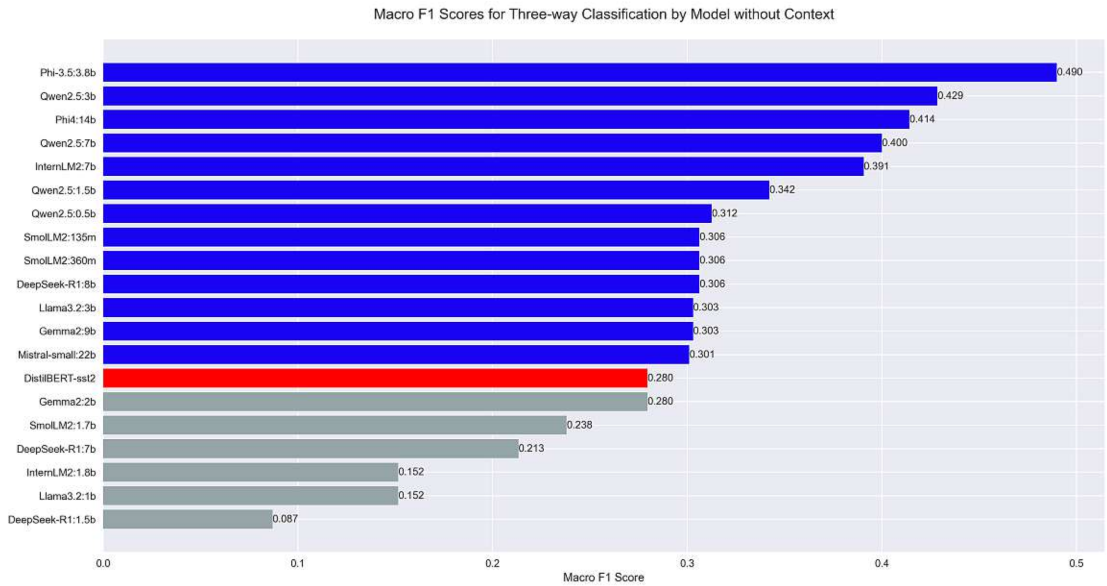

Figure 4.5: **Macro-F1 Scores for Three-way Classification in Zero-Context Evaluation.** Comparison of 20 language models' performance on classification tasks without contextual information (no transcript) using 20 YouTube comments. The y-axis represents the language models, while the x-axis shows the macro-F1 scores. Phi-3.5:3.8b achieves the highest score (0.490), followed by Qwen2.5:3b, Qwen2.5:7b, and Phi-4:14b. Smaller models such as DeepSeek-R1:1.5b and Llama3.2:1b show significantly lower performance. Blue bars highlight models which perform better than DistilBERT-sst2, which is highlighted in red.

4.9 Context-Enhanced Three-way Classification

Building on the previous phase, the next evaluation included the video transcript as context while retaining the three-way classification system. This phase aimed to assess whether the addition of contextual information led to improved classification results that more closely resembled human judgments. The prompt was enhanced with the transcript as *{context}*, as shown in Listing 4.9, while the response schema remained unchanged.

Listing 4.9: Prompt for three-way classification with context.

```
prompt = f"""
Task: Analyze the sentiment of the YouTube comment and provide
    detailed reasoning.
Video Transcript Context: {context}

Instructions:
    1. Analyze the comment's emotional tone, word choice, and
       context
    2. If transcript is provided, consider the video context in
       your analysis
```

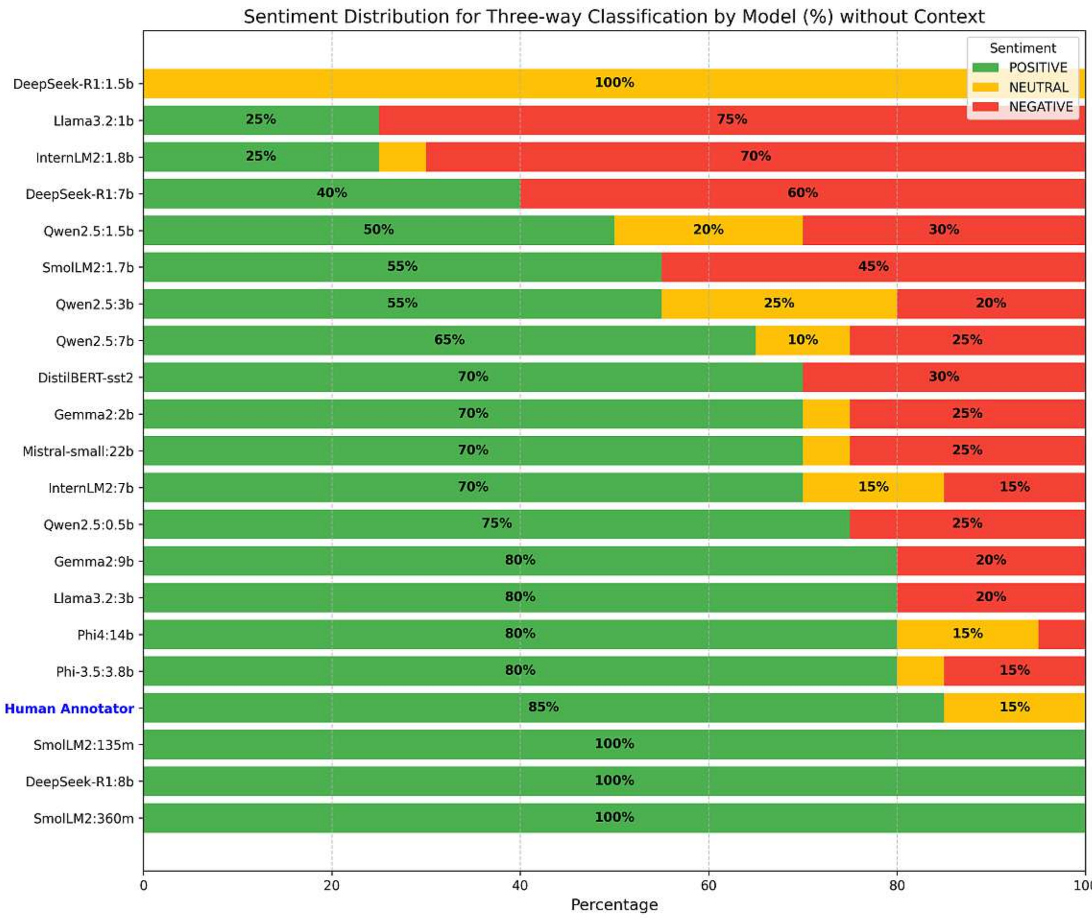

Figure 4.6: **Sentiment Distribution for Three-way Classification by Model (%) in Zero-Context Evaluation.** Comparison of 20 language models' performance on sentiment classification tasks without additional transcript, compared to human annotators' classifications. The y-axis represents the models, while the x-axis shows the percentage of comments classified as positive, neutral, or negative. The chart highlights variations in sentiment classification, with some models showing a strong bias toward positive sentiment (e.g., SmolLM2:135m, SmolLM2:360m, DeepSeek-R1:8b), others leaning toward negative sentiment (e.g., Llama3.2:1b, InternLM2.5:1.8b), and others provide a sentiment analysis closer to human annotations (e.g., Phi-3.5:3.8b, Phi-4:14b).

```
    3. Classify the sentiment as POSITIVE, NEUTRAL, or NEGATIVE
    4. Provide clear reasoning that includes:
        - Tone analysis
        - Key phrases or words that influenced your decision
        - Context consideration from the transcript if available
    5. Format your response as a JSON object with two fields:
        - sentiment: Your classification (POSITIVE/NEUTRAL/
          NEGATIVE)
        - reasoning: Your detailed analysis

YouTube Comment to analyze:
"{comment}"
"""
```

When examining the data in Figure 4.7, it can be seen that only four models achieve a macro-F1 score above 0.40 (Qwen2.5:7b, Phi-4:14b, Phi-3.5:3.8b, mistral-small:22b). However, the overall scores are higher and more aligned than they were *without context*. Similar to the binary classification evaluation *with context*, not all models benefited from the additional contextual cues.

Five models showed a decline in performance (DeepSeek-R1:8b, Phi-3.5:3.8b, Qwen2.5:0.5b, 1.5b, and 3b), while the following five models maintained their performance: DeepSeek-R1:7b, Llama3.2:1b, Llama3.2:3b, SmolLM2:135m, and SmolLM2:360m. Nine models showed an improvement, with mistral-small:22b not only being the best-performing model but also demonstrating the largest improvement, rising from a macro-F1 score of 0.301 in the *zero-context* evaluation to 0.535 in the context-enhanced evaluation.

In comparison, DistilBERT-sst2 now ranks 14th out of 20. Overall, more than 65% of the evaluated SLMs classified comments closer to the human ground truth than DistilBERT-sst2 did.

When analyzing the sentiment distribution in Figure 4.8, it becomes clear that the introduction of additional context influenced models that previously demonstrated consistently poor results. For example, DeepSeek-R1:1.5b, which had classified all comments as neutral in the prior *zero-context* evaluation and all as negative in both binary classifications, now classifies one comment as positive, two comments as negative, and the remaining 17 comments as neutral. This suggests that certain pieces of contextual information, combined with the slightly adjusted prompt for three-way classification, have contributed to changes in the model's behavior.

Similarly, InternLM2.5:1.8b, which in the prior *zero-context* evaluation had classified five comments as positive, one as neutral, and 14 as negative, now classifies the same five comments as positive, five as neutral, and only 10 as negative. While the model still ranks 17th out of 20 with a macro-F1 score of 0.235, the data indicates that the addition of context can influence model performance.

Overall, the best-performing model for three-way classification with contextual data is mistral-small:22b, which achieved a macro-F1 score of 0.535. Regarding that model, the

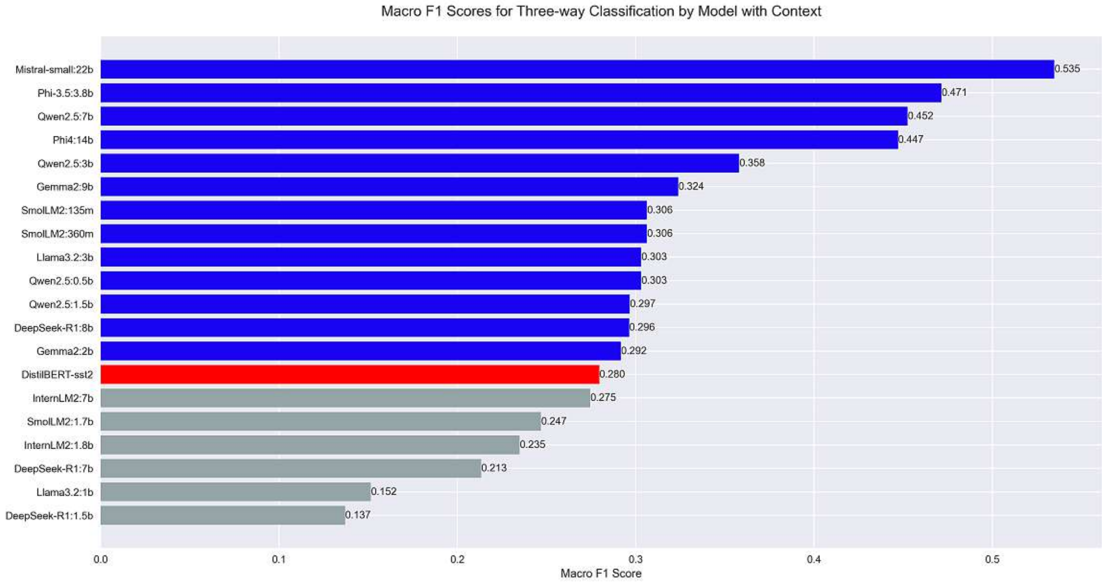

Figure 4.7: **Macro-F1 Scores for Three-way Classification with Context Evaluation.** Comparison of 20 language models' performance on classification tasks with contextual information, the video transcript, on 20 YouTube comments. The y-axis shows the language models and the x-axis shows the macro-F1 score. Mistral-small:22b achieves the highest score (0.535), followed by Phi-3.5:3.8b (0.471), while other models like DeepSeek-R1:1.5b and Llama3.2:1b show significantly lower performance. Blue bars highlight models which perform better than DistilBERT-sst2, which is highlighted in red.

data shows that contextual information in three-way classification improves alignment with human annotators (from macro-F1 score 0.301 to 0.535). Although the overall alignment of the models is closer to human annotators, not all models profited equally from the provided context (e.g., InternLM:7b from 0.391 in *zero-context* evaluation to 0.275 in *with context* evaluation).

Building upon the analysis of model performance, the next section provides an in-depth analysis of the computational efficiency and runtime performance of the top-performing models, with and without contextual data.

4.10 Findings: Performance Metrics Summary

This section analyzes the performance metrics of the best-performing models (Qwen2.5:3b, Qwen2.5:7b, Phi-4:14b, Phi-3.5:3.8b, and mistral-small:22b) based on macro-F1 score from the three-way classification, offering insights into the computational efficiency and runtime performance of the models. For clarity, the data in Tables 4.6 and 4.7 has been converted from nanoseconds to seconds.

The data reveals that the time required to load a model into memory *Initial Load Duration*

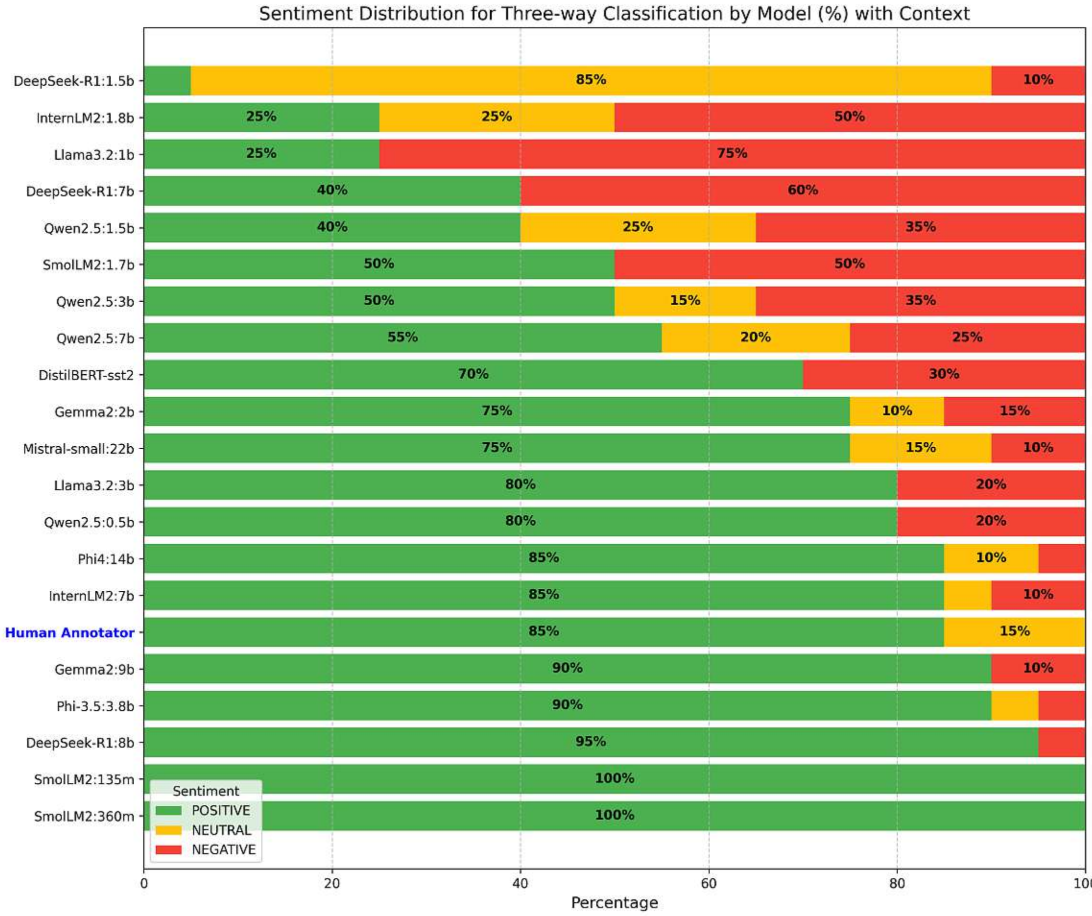

Figure 4.8: **Sentiment Distribution for Three-way Classification by Model (%) with Context Evaluation.** Comparison of 20 language models' performance on sentiment classification, including an additional transcript, with human annotators. The y-axis represents the models, while the x-axis shows the percentage of comments classified as positive, neutral, or negative. The chart highlights variations in sentiment classification, with some models showing a strong bias toward positive sentiment (e.g., SmolLM2:135m, SmolLM2:360m), others leaning toward negativity (e.g., Llama3.2:1b), and some providing a more balanced sentiment analysis (e.g., Mistral-small:22b, Qwen2.5:7b).

Model	Initial Load Duration	1. Prompt Eval Duration	2. Prompt Eval Duration	3. Prompt Eval Duration
Qwen2.5:7b	0.56	3.68	0.60	0.14
Qwen2.5:3b	0.57	1.74	0.29	0.07
Phi-4	0.79	7.79	1.26	0.24
Phi-3.5	1.31	2.83	0.33	0.07
mistral-small:22b	0.52	10.31	1.74	0.35

Table 4.5: Time in seconds for each model to load into memory and time it takes for prompt evaluation of the first 3 prompts

is lower than the time it takes to evaluate the first prompt *1. Prompt Eval Duration* where the majority of time is spent. The prompt evaluation duration significantly decreases for all models by the third run and remains consistently low thereafter. Table 4.5 presents the durations for loading the first comment and evaluating the first, second and third prompt for each model. These values are system-dependent and may not be reproducible; they serve as an indication of the models loading speed and internal tokenization time. The evaluation was conducted on a Mac Mini M2 Pro with 32 GB of RAM and 19 GPU cores.

Next, the *Performance Metrics Summary* for the best performing models in the three-way classification *without context* is presented.

4.10.1 Evaluation without context

The *Performance Metrics Summary* of the best performing models in the three-way classification *without context* (Qwen2.5:3b and 7b, Phi-4:14b, Phi-3.5:3.8b) is presented in Table 4.6. This includes the mean total-, load-, prompt-, and evaluation durations, as well as the prompt evaluation count for analyzing a comment with an SLM.

Phi-4:14b stands out with a mean *Total Duration* of 13.61 seconds for analyzing a comment, with most of the time spent on *Eval Duration*, which refers to the time the model takes to generate output tokens. The length of the response is also the longest in Phi-4:14b, with an *Eval Count* of 219.75, while Qwen2.5:7b has the shortest response (*Eval Count* = 133.15). The data suggests that smaller models tend to deliver faster response times, with Qwen2.5:3b being the fastest and smallest, followed by Phi-3.5:3.8b, Qwen2.5:7b, and Phi-4:14b.

Following the evaluation *without context*, the next section examines the *Performance Metrics Summary* for the best performing models in three-way classification *with context*.

4.10.2 Evaluation with Context

The *Performance Metrics Summary* for three-way classification *with context*, presented in Table 4.7, reveals notable differences among the best-performing models (Qwen2.5:7b, Phi-

Model	Total Duration	Load Duration	Prompt Eval Count	Prompt Eval Duration	Eval Count	Eval Duration
Qwen2.5:7b	4.47	0.01	204.30	0.31	133.15	4.15
Qwen2.5:3b	2.75	0.01	204.30	0.15	154.80	2.59
Phi-4:14b	13.61	0.01	185.55	0.60	219.75	13.00
Phi-3.5:3.8b	3.64	0.01	214.55	0.16	190.55	3.47

Table 4.6: Performance metrics summary (mean) for processing a comment. *Duration* in seconds. *Count* in tokens.

Model	Total Duration	Load Duration	Prompt Eval Count	Prompt Eval Duration	Eval Count	Eval Duration
Qwen2.5:7b	5.09	0.01	644.30	0.38	145.65	4.71
Phi-4:14b	13.96	0.01	617.55	0.73	214.05	13.22
Phi-3.5:3.8b	5.08	0.01	705.55	0.23	237.40	4.84
mistral-small:22b	11.54	0.01	689.25	1.04	128.15	10.48

Table 4.7: Performance metrics summary (mean) for processing a comment with transcript as context. *Duration* in seconds. *Count* in tokens.

4:14b, Phi-3.5:3.8b, and mistral-small:22b). An increase in *Total Duration* is observed for the models Qwen2.5:7b, Phi-4:14b, and Phi-3.5:3.8b, with Phi-3.5:3.8b showing the most significant increase, from 3.64 to 5.08 seconds, and Phi-4:14b showing the least, increasing from 13.61 to 13.96 seconds. In general, the rise in *Total Duration* is expected due to the additional context that has to be processed, as also reflected in the higher *Prompt Eval Count*. Notably, for the models Phi-3.5:3.8b and Qwen2.5:7b, the incorporation of context resulted in a slight increase in the *Eval Count* (237.40 and 145.65 *with context*, respectively, compared to 219.75 and 133.15 in *zero-context*), leading to a corresponding increase in *Eval Duration*. Interestingly, despite being the largest model in terms of parameter size, mistral-small:22b demonstrated a lower *Total Duration* than the smaller Phi-4:14b and produced the shortest response, as indicated by its *Eval Count* of 128.15.

4.11 Supplementary Analysis

An additional exploratory investigation was conducted using a multi-tier classification system with SLMs. Although this approach provided a more granular sentiment analysis, it was excluded from the comparative analysis due to participant feedback, which indicated cognitive overload during the user evaluation study. Furthermore, no comparable model data was available for inclusion.

CHAPTER 5

User Research

This chapter presents the findings from both the user preference study and the user evaluation study in a structured manner. It first examines insights from the user preference study and the low-fidelity prototypes, discussing how they informed the development of a functional, high-fidelity prototype. Then it outlines the pilot studies conducted prior to the user evaluation, highlighting key challenges and pitfalls encountered. Based on insights from the pilot testing, the final tools (Tool A, Tool B, and Tool C) for the user evaluation study were developed. Finally, the findings from the user evaluation are presented across four key areas: (1) pre-task questionnaire insights, which capture user expectations and initial perceptions; (2) user-task observations, which document key interactions and behavioral patterns during the tasks; (3) post-task questionnaire results, which reflect user feedback and experience assessment; and (4) thematic analysis findings, which identify key themes emerging from the collected qualitative data.

5.1 User Preference Study

The project began with creating low-fidelity prototypes using Figma to explore different ways of displaying sentiment analysis results within the YouTube comment section. These low-fidelity prototypes (see Figure: 5.1) included four different design approaches: Prototype 1, where negative comments were hidden with a disclaimer; Prototype 2, where negative comments were fully redacted while positive ones were highlighted; Prototype 3, which partially redacted negative words; and Prototype 4, where comments were categorized into positive, negative, and neutral, accessible via a drop-down menu. In Figure 5.1, Prototype 4 includes two representations to better illustrate how the drop-down menu worked for categorizing comments.

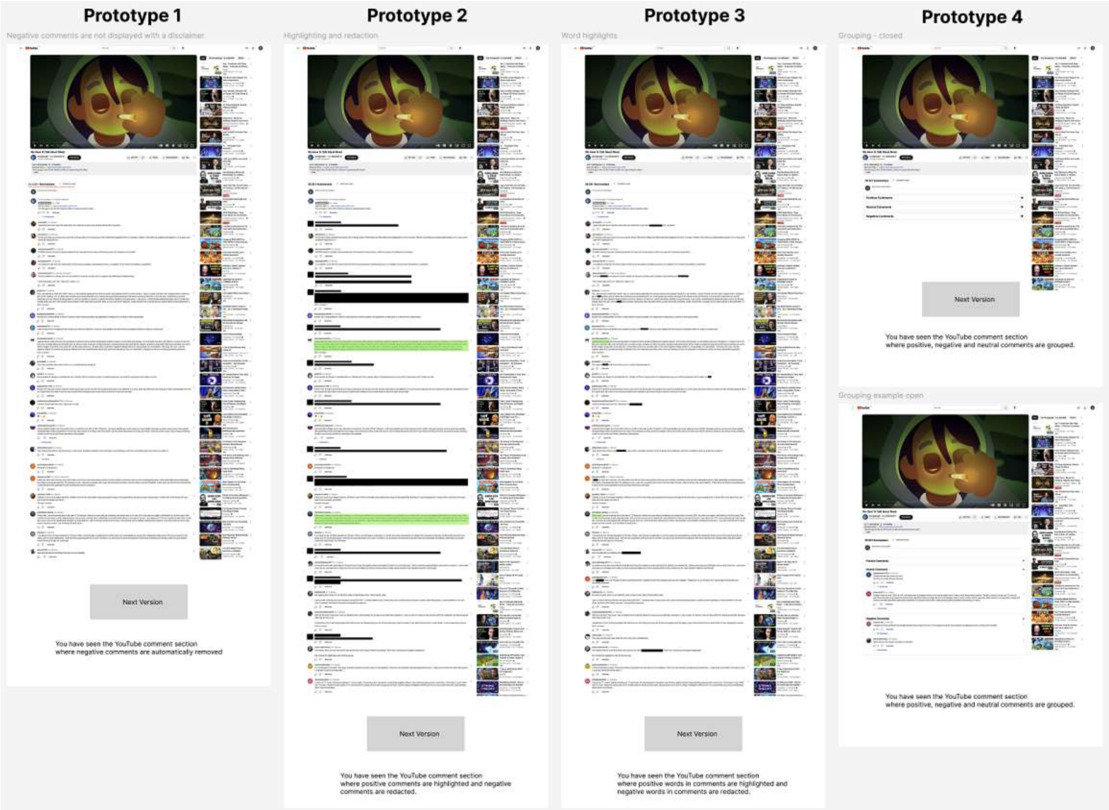

Figure 5.1: Interactive low-fidelity Figma prototypes used in the user preference study, showcasing different sentiment classification design approaches. Prototype 1: Negative comments were hidden, and a disclaimer informed users about the filtering next to the total comment count. Prototype 2: Negative comments were fully redacted, while positive ones were highlighted. Prototype 3: Negative words in comments were partially redacted. Prototype 4: Comments were initially hidden and categorized into positive, negative, and neutral, accessible via a drop-down menu.

5.1.1 User Preference Study: Findings

To evaluate what kind of tool was needed and preferred by users for sentiment analysis of YouTube comments, a user preference study was conducted. Participants provided feedback on four differently designed low-fidelity prototypes, as shown in Figure: 5.1. The following sections present the findings based on transcripts and notes taken during the interviews (see Appendix: 7), which later informed the design of high-fidelity prototypes implemented as browser extensions (further described in Section: 5.2). These prototypes ultimately led to the development of the final tools for the user evaluation study.

Perceived Usefulness and Community Context

Participants had varying opinions on the usefulness of a comment analysis tool in general, as four out of six stated that they rarely actively read or engage with comments. P4 highlighted the gaming community as particularly prone to unhelpful or even hateful comments, leading them to avoid reading them. Others (P2, P3) pointed out that polarizing topics, such as game reviews or controversial initiatives (e.g., "Nintendo should be boycotted"), often attract negativity. P6 additionally noted that community dynamics in comment sections vary significantly across different content categories (e.g., news, reviews, music, lifestyle, education) and suggested that a comment analysis tool should adapt its filtering based on the specific community context.

Comment Engagement and Consumption Behavior

Participants reported engaging with YouTube comments for various reasons, though overall, they rarely actively interact with them. Several participants (P3, P4, P5) primarily use the comments to find timestamps referencing specific information, such as song names or key moments in a video. Others (P1, P2) browse the comments to seek clarification on video content, looking for answers to shared questions or to confirm their own observations.

The decision to read comments often depends on the video type. One participant (P6) engages with comments on modeling channels to find constructive feedback on outfits but avoids them in self-development videos, perceiving them as lacking valuable input. In the context of political content, participants (P1, P3, P4) generally view the comments as predominantly negative. Additionally, two participants (P5, P6) noted that smaller YouTube channels tend to receive more positive and encouraging feedback, though this is not their primary interest.

Participants also reported being more likely to read comments when prompted by the video itself (e.g., "share your view on XYZ in the comments"), in tech reviews for product feedback, or for entertainment purposes (e.g., humorous comments). Regarding comment-writing behavior, two participants (P4, P6) stated that they only comment when they feel gratitude or have constructive feedback to share. Another participant (P2) emphasized that the comment section primarily serves as a communication channel between viewers and creators rather than among viewers.

Attitudes Toward Negative Comments

None of the participants reported flagging or responding to negative comments. However, P6 suggested that some content creators deliberately provoke controversy to exploit the YouTube algorithm and drive engagement through comment activity.

Regarding comment filtering, participants generally agreed that insults should be removed, while general negative opinions should remain—though they did not provide concrete reasons for this distinction. Some participants (P1, P4) expressed interest in filtering out repetitive spam, such as “first” comments (a comment from a viewer who indicates being the first to comment). Although they believed that some users might prefer negative comments to be hidden, all participants preferred them to remain visible rather than hidden.

Concerns and Filtering Preferences

P1, P2, P5, and P6 voiced concerns about automated filtering, citing fears of censorship, misclassification, and over-filtering. While some participants (P3, P4) viewed filtering as a potential “quality of life” feature—allowing users to avoid negativity or redundancy in comments if desired—others questioned its necessity, arguing that comments were not important enough to justify a dedicated tool. Additionally, some expressed frustration over the growing tendency of news channels to disable comments altogether.

Following the analysis of qualitative feedback from the user preference study, the next section discusses participants’ responses to the low-fidelity prototypes themselves.

5.1.2 User Reactions to Low-fidelity Prototype Designs

This section examines participants’ responses to the low-fidelity prototypes, focusing on how different display methods influenced their perception and engagement. These initial prototypes served as a foundation for evaluating interaction and visualization methods, despite lacking a fully developed interface. However, the study revealed that none of the proposed display methods were perceived as engaging or useful. A key insight was that participants strongly valued transparency—no content should be hidden or redacted.

In general, participants found that Prototype 2 disrupted the reading flow too much, while Prototype 3 encouraged them to guess the meaning of partially redacted words. Prototype 4 faced strong criticism for preemptively categorizing comments as positive, negative, or neutral, limiting users’ ability to form independent opinions.

With Prototype 1, which simply displayed a message stating that *52 comments were hidden due to negative sentiment*, all participants expressed a desire to see which comments had been hidden. When asked whether this would still apply if the message read, *2000 comments hidden* all participants responded that they would likely check the first few hidden comments to verify the tool’s accuracy and would want the option to display them again.

Despite some minor positive feedback, the overall discussion led to the rejection of all initial design concepts in favor of a simpler, less obtrusive approach—one that preserves the original order and layout of the comment section without hiding any content.

5.1.3 Implications for the High-Fidelity Prototypes

Participants emphasized the need for transparency in filtering decisions, aiming to balance critical discussions with the removal of harmful content. While opinions on comment engagement varied, the findings suggest that an effective analysis tool should offer customizable classification and filtering options, giving users control and clarity. This would help ensure the tool's acceptance across different types of "YouTube comment communities", which can differ based on the video content and audience.

While these insights from the user preference study provide valuable direction, pilot testing is essential to confirm that the tool's features function as intended in real-world settings. Pilot testing is not aimed at gathering a representative sample of users, but at collecting feedback to refine both the tool and study methods [14]. The results from the pilot studies are used to reconsider aspects of the tool and study design [116]. This step is necessary because what users want may not always align with how the tool works in practice, which is why pilot testing plays a crucial role in ensuring that the tool meets real-world needs.

Based on the insights from the user preference study, the high-fidelity prototypes were developed specifically for pilot testing. These prototypes, now functional, were tested with users to gather feedback that informed refinements to the final tool, which will be discussed in Section 5.3. The details of the high-fidelity prototypes and the findings from the pilot tests are covered in the following section.

5.2 High-fidelity Prototypes

Initially, based on feedback from the participants of the user preference study, two high-fidelity prototypes (see Figure: 5.2) were developed and evaluated in a pilot. These prototypes incorporated an approach that allowed users to create their own classifications based on their needs and preferences. Details of the design approach for High-Fidelity Prototypes 1 and 2 can be found in Section 5.2.2.

Following the results of the first pilot, two additional high-fidelity prototypes were developed. One of these prototypes implemented "pre-defined classifications", while the other was represented through an external website (see Figure: 5.3). The design approaches of these new prototypes are discussed in Section 5.2.4, and they were tested in a subsequent pilot.

An initial step in the process was selecting a suitable platform for prototype development. The participants' usage behavior played a key role in this decision, and the rationale behind the selection of Google Chrome will be outlined in the next section.

5.2.1 Choice of Google Chrome as Target Platform

Since the high-fidelity prototypes needed to be functional extensions, a decision had to be made regarding the browser for which the extension should be developed. When asked where they consume YouTube content, most participants distinguished between usage on the go versus at home or work. The majority accessed YouTube via the mobile app on their smartphones (n=5), while others used web browsers. Among these, Google Chrome (n=3) was the most frequently mentioned, followed by Firefox (n=2), Opera (n=1), and Brave Browser (n=1). One participant (P1) also reported using a Smart TV to watch YouTube.

Given that this thesis aims to develop a browser extension, Google Chrome was selected as the target platform, as it was the most commonly mentioned browser among participants. Furthermore, according to [109], Google Chrome holds a global market share of 66.29% as of February 2025, making it the most widely used browser across all platforms (desktop, tablet, and mobile). It is followed by Safari (18.01%) and Microsoft Edge (5.33%), while Firefox and Opera account for 2.63% and 2.09%, respectively.

A closer look at Austria's browser usage statistics [109] reveals a similar trend, with Google Chrome leading at 48.59%, followed by Safari (23.08%) and Microsoft Edge (10.18%). Firefox holds a market share of 9.42%, while Opera accounts for 3.54%. These numbers further support the decision to prioritize Google Chrome as the primary platform for the browser extension developed in this thesis.

In addition to being the most commonly used browser globally, choosing Google Chrome allows the extension to reach the largest possible audience. As the browser with the highest market share, Chrome provides the opportunity to engage a broad user base, maximizing the extension's accessibility and potential impact.

5.2.2 Design High-fidelity Prototypes 1 and 2

The first two high-fidelity prototypes, developed as Chrome extensions, are shown in Figure: 5.2. These prototypes were designed to provide users with advanced natural language processing models to classify YouTube comments.

The first prototype, titled "YouTube Comment Analyzer" uses facebook's BART model (*bart-large-mnli)* [55] in the backend and allows users to define their own classes for comment categorization. After defining these classes, users can choose how the results are displayed, such as by highlighting or labeling comments that match the defined categories.

Once the user presses "Analyze Comments" the BART model processes the comments and assigns them a class with an associated confidence score. For example, a comment might be labeled as "informative – confidence score: 90%". Additionally, users can filter comments based on the predefined classes to adapt the comment section to their needs.

The second high-fidelity prototype, on the right in Figure 5.2 "SLM Comment Analyzer including custom classes and context awareness (transcript)", differed in that it allowed users to specify a class and provide examples in natural language that represented that class. These examples were considered during the analysis of the comments by the SLM. For instance, a user could define a category called "not interesting" and provide examples like "comments praising the video or creator", "comments which are not directly related to the video" directly reflecting the proposed approach from the user preference study.

Additionally, this version was capable of referencing the video transcript as context during the analysis, providing explanations for its sentiment analysis classifications. The filter and display options were the same as those in the previously described prototype.

To evaluate the design of these high-fidelity prototypes, a pilot study (Pilot 1) was conducted, which will be described next.

5.2.3 Pilot 1

Pilot 1 evaluated two high-fidelity prototypes (Chrome extensions, see Figure: 5.2), inspired by the findings from the user preference study. These prototypes allowed users to define their own classification criteria and provide examples, utilizing facebook's BART model [55] and an SLM in the background.

During the pilot test (n=2), both participants struggled to define meaningful classification categories, finding it difficult to conceptualize categories that were relevant to their needs. Additionally, providing examples for categories for the SLM-based approach proved challenging, complicating the classification process and hindering the overall user experience. As a result, the custom class and example features were excluded from the final user evaluation study.

Given these challenges, it was necessary to step back and reconsider the design of the prototypes. The feedback from pilot study 1 revealed that the high-fidelity prototypes did not fully meet user needs, prompting the development of new prototypes that better aligned with users' needs and expectations.

5.2.4 Design High-fidelity Prototypes 3 and 4

Building on insights from the literature review [87, 80] and findings from pilot study 1, an alternative approach to sentiment analysis tools was explored. Previous research often evaluated tools that operated on external websites rather than directly within YouTube [87, 80]. To compare these approaches, two new high-fidelity prototypes were developed: one as a Chrome extension with enhanced features based on user feedback (Prototype 3) and another one as a traditional web-based tool running a BERT-based model in the background (Prototype 4).

The Chrome extension was specifically designed to classify comments as positive, negative, or neutral. This version incorporated display options from the user preference study, a summarization feature, and the well-received labeling feature from High-Fidelity Prototypes 1 and 2. Additionally, the extension utilized the video's transcript as context and was capable of explaining its reasoning behind classifications.

A key improvement in this iteration was the introduction of user-selectable SLMs for classification within the extension. The goal was to identify the preferred model, which could help determine the best-performing one. To prevent bias toward well-known models, the model names were anonymized and labeled as “Model 1”, “Model 2”, all the way up to “Model 6”.

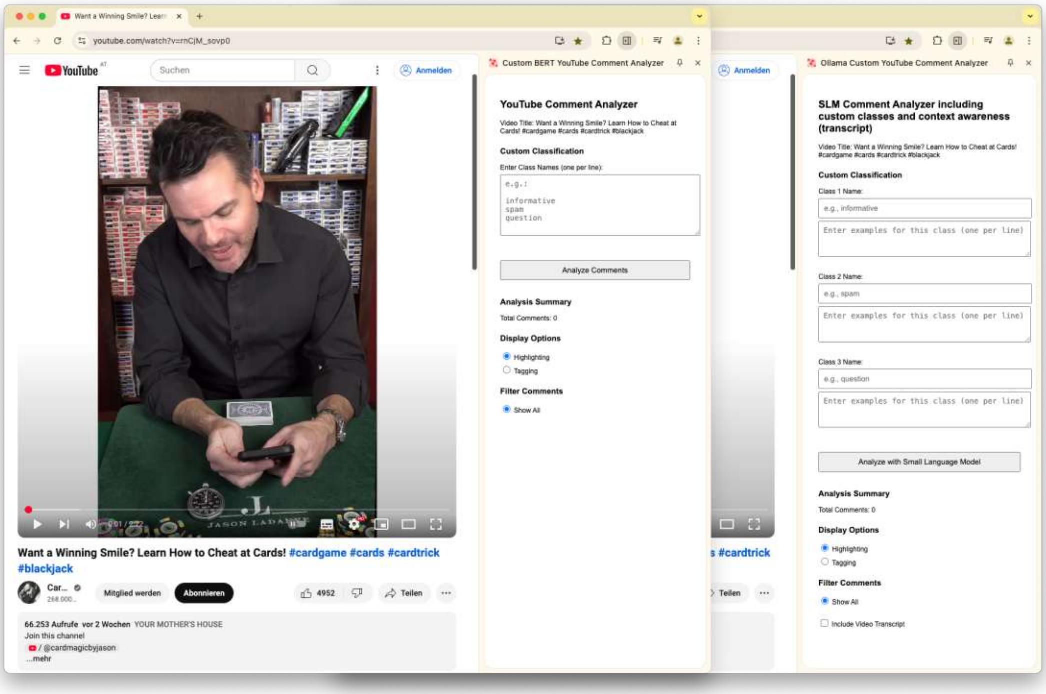

Figure 5.2: Left: user can specify their own classes with the BART model (*bart-large-mnli*). Right: user can specify own classes and provide examples with the SLM. Approval to use Jasons face and channel name was obtained (see Appendix 7).

To evaluate their effectiveness and usability, the two new Prototypes (3 & 4) were tested in a second pilot study.

5.2.5 Pilot 2

Participants (n=2) were tasked with comparing the results—both sentiment labels and reasoning—generated by multiple models for each comment. These results were also compared to those produced by the dedicated website.

A detailed study protocol was prepared, including an 11-minute and 43-second training video [1], along with printed questionnaires and task sheets.

In this iteration, the custom classification feature was removed entirely. Instead, a dedicated website for sentiment classification using DistilBERT-sst2 was introduced to facilitate the planned comparison between a browser extension running directly on YouTube and an external tool. The hypothesis was that users would prefer the browser extension for its seamless integration into YouTube over the dedicated website.

[1] Link to the training video: `https://www.youtube.com/watch?v=9sdYVs2IRAY` - last accessed 19.03.2025

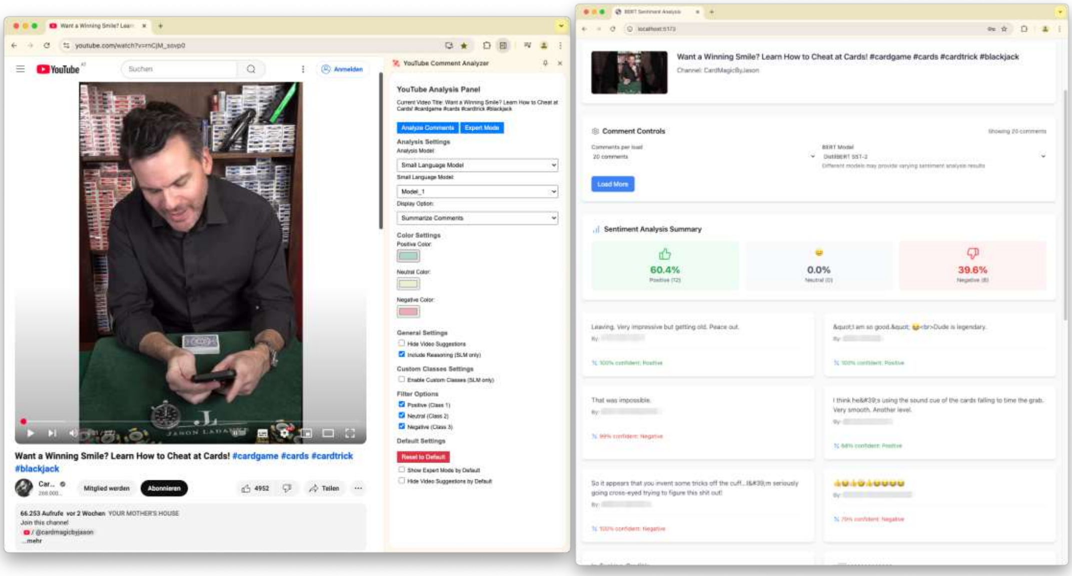

Figure 5.3: Left: User can specify the model and choose from varying display methods: summarization, grouping, highlighting, labeling. Right: User can navigate to a dedicated website for sentiment analysis with a BERT-based model.

However, during the pilot study, it became evident that participants were evaluating the tools primarily based on their visual design rather than functionality. The external website, with its more polished interface, was consistently favored, as noted in participant feedback. This introduced an unintended bias, leading to the decision to abandon the direct comparison between the extension and the website for the user evaluation study. Instead, the focus shifted to assessing the usability and effectiveness of the Chrome extensions alone.

The cognitive load of the study proved too high, and the sessions became overly time-intensive, with one being canceled after two hours due to participant fatigue. Issues included the lengthy training video, an overwhelming number of options to compare (summarization, grouping, highlighting, filtering), and the challenge of choosing from six models. While the YouTube videos comment section was interesting, its 14-minute video length made it difficult for participants to retain focus.

For the user evaluation study, the decision to use a single extension with multiple sentiment display options was reconsidered and ultimately discarded due to concerns about complexity and the potential for errors. Instead, three separate extensions were used in the final implementation.

Several additional features, considered based on feedback from the user preference study and pilot tests, were excluded to maintain simplicity. These included: comment grouping and sorting, automatic pre-sorting, background highlighting, custom classification, custom colors for classes, a summarization feature, a dedicated website, and free model selection.

The insights gathered from pilot study 1 and 2, where participants provided feedback on both the study design and the design of the high-fidelity prototypes, were instrumental in refining the final Chrome extensions (Tool A, Tool B, Tool C) and the study design, which are discussed in the next section.

5.3 Final Chrome Extensions

Building on the feedback from pilot studies 1 and 2, the final set of Chrome browser extensions was developed. These extensions were designed to test different variations of sentiment analysis and their impact on user experience within the context of YouTube comments. The tools were developed with specific design choices in mind, guided by participants' feedback on the usability of the high-fidelity prototypes. To allow for a systematic comparison, the tools were presented in a specific sequence, facilitating a gradual introduction of features that would help participants evaluate the evolution of the sentiment analysis methods (see Figure: 5.4). Each extension has two checkboxes at the bottom of the sidepanel which were used to enable the cinema mode of the video and to hide suggested videos during the study. The purpose was to minimize potential distraction (e.g., catchy thumbnails or titles from other videos) for participants during the study and to achieve a consistent user interface setup during all sessions.

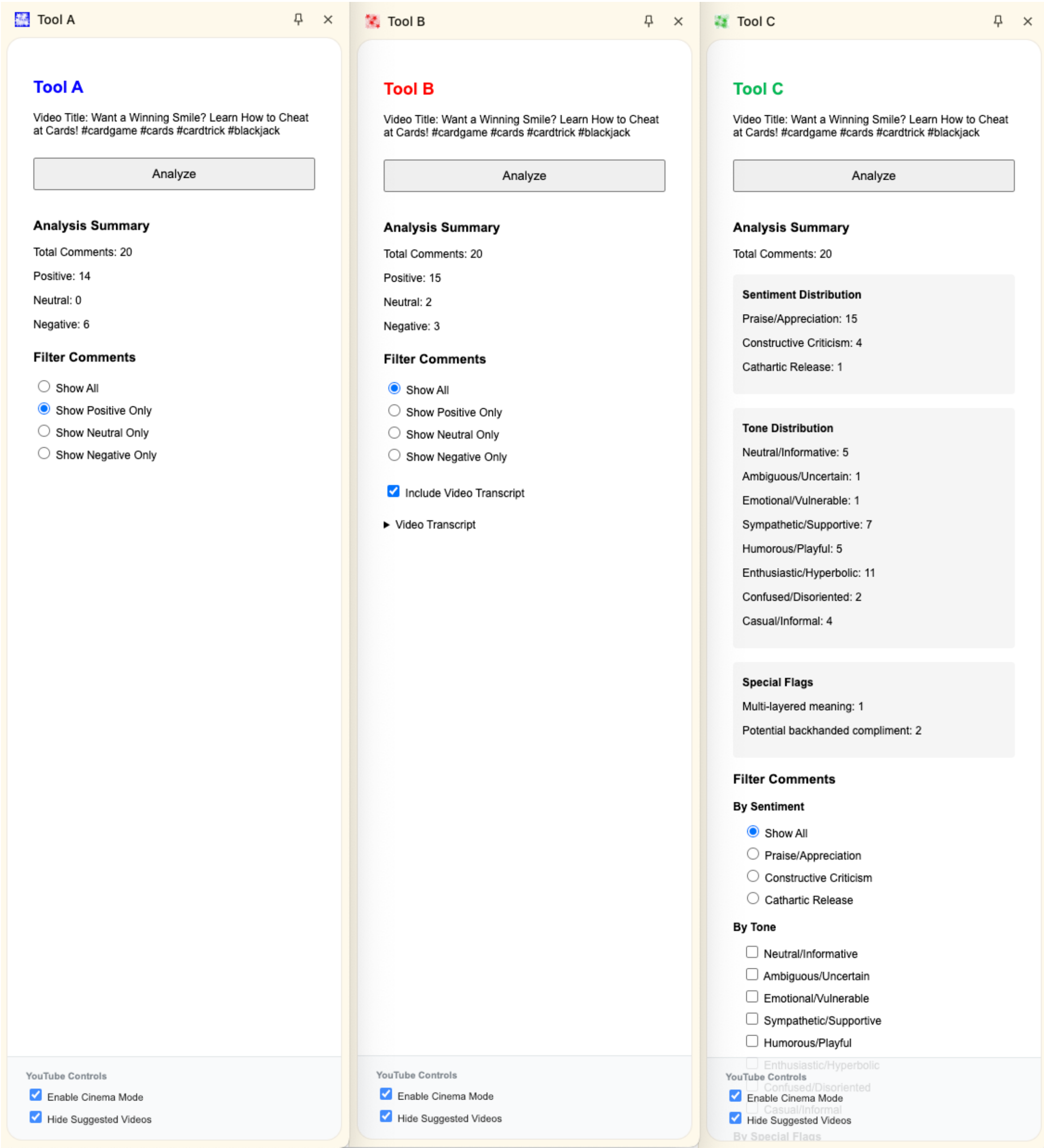

Figure 5.4: The UI of the sidepanels displayed side-by-side for better comparison. The participants will go through one after the other.

Tool A (see Figure: 5.5) represented the standard approach to sentiment analysis. It provided confidence scores for positive and negative classifications but lacked any explanatory features. It was used to establish a baseline for traditional sentiment analysis methods. It has to be noted that visually the "neutral" option was shown in the sidepanel even though the model used was only capable of doing binary classification. The intention was to streamline the comparability of Tool A and Tool B.

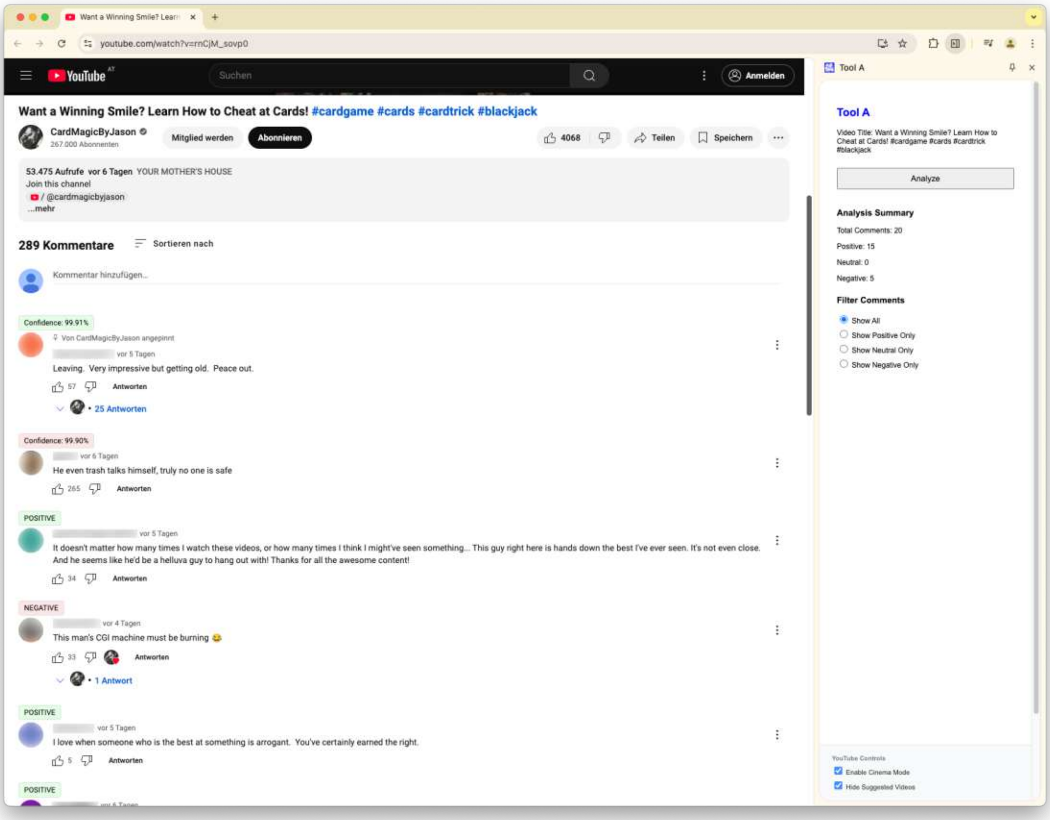

Figure 5.5: Tool A: On the left the YouTube comment section is displayed showing comments of viewers which got "labeled" by Tool A as either positive or negative. Clicking on the label reveals the confidence-score of the model for the classification. On the right, the sidepanel of the Chrome extension is displayed, offering an "Anlayze" button (to start the analysis), a summary of the results and filter options.

Tool B (see Figure: 5.6) offered the classification categories positive, negative, neutral but replaced the confidence scores of Tool A with explicit reasoning for each classification. This added complexity, enabled a direct comparison with Tool A, while introducing participants to the concept of "model-reasoning" for classifications.

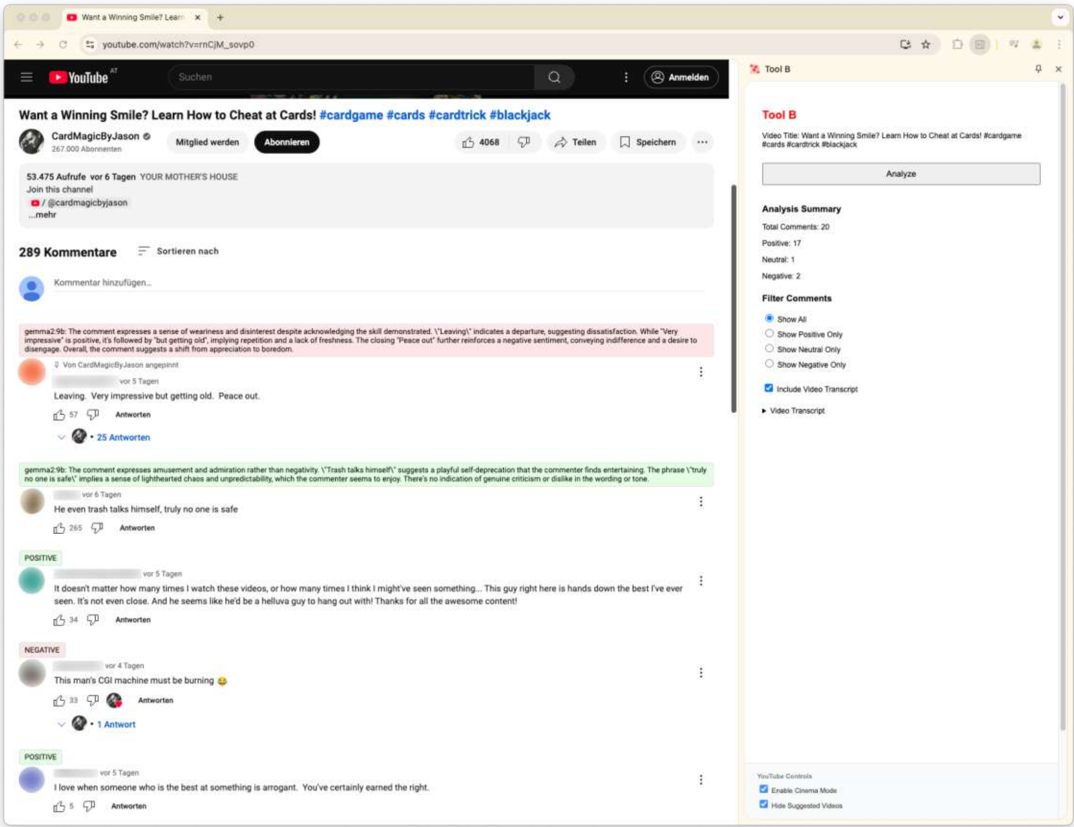

Figure 5.6: Tool B: On the left the YouTube comment section is displayed showing comments of viewers which got "labeled" by Tool B as either positive, negative or neutral. Clicking on the label reveals an explanation of the model for the chosen classification. On the right, the sidepanel of the Chrome extension is displayed, offering an "Anlayze" button (to start the analysis), a summary of the results and filter options as well as a checkbox to indicate if the transcript is used during classification.

Tool C (see Figure: 5.7), presented last, expanded the classification categories beyond the simple positive, negative, neutral classes, while still maintaining the reasoning component introduced by Tool B. By presenting the tools in this sequence, participants were able to progressively deepen their understanding and make more meaningful comparisons between the different approaches.

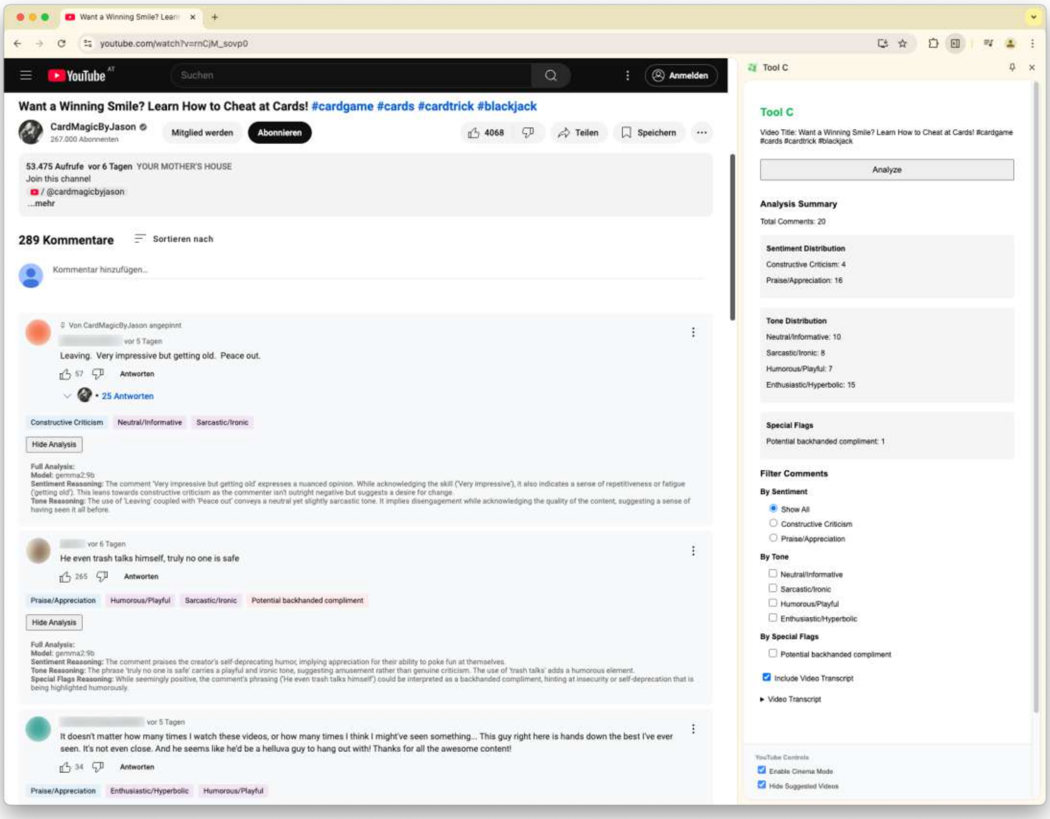

Figure 5.7: Tool C: On the left the YouTube comment section is displayed showing comments of viewers which got "labeled" by Tool C in a primary class (blue), secondary classes to describe the tone (purple) and a third class to indicate special flags (orange). Clicking on the label or the "Show Analysis" button reveals an explanation of the model for the chosen classification. On the right, the sidepanel of the Chrome extension is displayed, offering an "Anlayze" button (to start the analysis), a summary of the results and filter options as well as a checkbox to indicate if the transcript is used during classification.

5.4 User Evaluation Study

This section outlines the user evaluation study, which aimed to assess the usability and effectiveness of the final Chrome extensions. The study included the following components: a pre-task questionnaire, a training video, user tasks, a post-task questionnaire, and an audio-recorded semi-structured interview. The pre-task questionnaire (see Appendix: 7) collected baseline information on participants' experience with sentiment analysis tools and Chrome extensions. The training video provided an introduction to the tools and guided participants on how to use them during the study.

Participants were then asked to complete specific tasks (see Appendix: 7) with the Chrome extensions, which were followed by a post-task questionnaire (see Appendix: 7) to evaluate their experience. Finally, a semi-structured interview (see Appendix: 7) was conducted to gain deeper insights into participants thoughts, challenges, and preferences. The interviews were audio-recorded, and the transcripts, along with notes, were used for the thematic analysis. This analysis helped to identify key themes and insights about the Chrome extensions.

5.4.1 Pre-Task Survey: Analysis

Before interacting with the sentiment analysis tools, participants completed a pre-task survey to establish baseline insights into their YouTube usage patterns and attitudes toward automated sentiment analysis. The survey gathered demographic data, assessed comment-reading habits, and explored initial perceptions and potential concerns about such tools. The following sections present the analysis of the collected data.

YouTube Usage Patterns

The analysis revealed that all participants (n=7) used YouTube primarily for entertainment. More than half (P1, P3*, P4*, P7) [2] also used it for information and news, while three participants (P1, P4*, P7) accessed educational content. Music consumption was common among most (P1, P3*, P4*, P5, P6*), and no additional usage patterns were mentioned.

Comment Reading Behavior and Motivation

Regarding comment engagement, four participants (P3*, P5, P6*, P7) reported rarely reading comments. P4* stated they "nearly always" read comments, P1 "sometimes," and P2 "never." Motivations for reading comments included gaining additional insights (P1, P4*) and entertainment (P4*). Notably, both P1 and P4* never engaged to respond to comments. Participants with less frequent engagement (P1, P3*, P6*, P7) mentioned perceived irrelevance or general disinterest. None reported negative past experiences with YouTube comments.

Perceived Helpfulness of a Comment Analysis Tool

Opinions on the usefulness of a comment analysis tool varied. Three participants (P1, P3*, P4*) thought it would be helpful, with P6* expressing strong interest, calling it "very helpful". P2 and P5 were neutral, while P7 thought it would not be helpful. No additional opinions were shared.

[2] Participants with an * (e.g., P3*) also took part in the user preference study

Key Concerns Regarding the Automated Analysis Tool

Participants raised several concerns regarding the potential drawbacks and ethical implications of an automated sentiment analysis tool for YouTube comments. These concerns are summarized below:

- **Over-filtering and removal of legitimate content:** Concerns were raised about over-filtering, where legitimate comments, such as valid criticisms, might be mistakenly flagged as inappropriate, limiting meaningful discussions (P1, P3*, P4*, P5).
- **Bias in data analysis and decision-making:** Participants were worried that the tool might reflect biases, misclassifying diverse opinions, especially in sensitive or controversial topics, leading to skewed results (P1, P4*, P5).
- **Censorship and free speech limitations:** Fears of automated censorship were expressed, particularly regarding political or ideological content, where moderation might suppress non-mainstream views (P4*).
- **Privacy risks, especially with private information:** Concerns were raised about the tool unintentionally processing private or sensitive information from YouTube comments, which could lead to privacy violations (P6*).
- **Need for transparency in decision-making:** The importance of transparency in how decisions are made by the tool was emphasized to ensure user trust, especially in content moderation scenarios (P7).

Neutral Perspectives and Additional Context

Two participants (P2, P5) expressed no concerns about the tool. P7 noted their limited engagement with comments, which led them to conclude that such a tool would not be helpful for them.

Summary of Findings (Pre-Task Survey)

The pre-task survey revealed diverse YouTube usage patterns, with all participants (n=7) using it for entertainment and varying uses for information, education, and music. Comment engagement was generally low, with most participants (P3*, P5, P6*, P7) rarely reading comments due to perceived irrelevance, similar to result from the user preference study. Those who did engage (P1, P4*) sought additional insights or entertainment but did not participate in discussions.

Regarding the proposed comment analysis tool, opinions were mixed. Some (P1, P3*, P4*, P6*) saw value in filtering and analysis, while others (P2, P5, P7) were skeptical. Concerns focused on over-filtering, bias, free speech limitations, and privacy. Participants emphasized the need for transparency in content moderation. These findings suggest that while an automated comment analysis tool may appeal to some, it must balance filtering accuracy, user control, and transparency to ensure broader acceptance.

The next section will discuss how participants interacted with the tools and their impressions during the test.

5.4.2 User Tasks: Analysis

After the pre-task study, participants were asked to evaluate three different Chrome extensions (Tool A, Tool B, and Tool C) designed for sentiment analysis of YouTube comments. The tasks participants completed involved initial content exposure (watching a 2-minute YouTube video and reading 20 comments) followed by systematic tool analysis using each tool, where they expressed their agreement with classifications, evaluated explanations, and assessed tone analysis.

A "think-aloud" [33] method was employed, allowing participants to verbalize their thoughts while interacting with each tool. This method was originally introduced by Ericsson and Simons in 1993 [33] and offers a way to use verbalization in research. Combining think-aloud methods with interviews helps capturing participants' reactions more effectively [30]. Verbalizing general thoughts does not impact accuracy, but explaining reasoning can. Concurrent reports are richer because they rely on immediate cues, while retrospective reports are less detailed due to memory limitations [33].

Because of the valuable insights and the focus on participants' reactions during their evaluation processes, this method was employed in this thesis. It created the opportunity for the participants to verbalize specific points of praise and criticism in a structured way. The following section is a compilation and summarization of the notes, the researcher took during this process and it highlights the key aspects, articulated by the participants.

Several participants (P1, P2, P5, P6*) criticized Tool A's inability to accurately interpret ironic or sarcastic comments, as well as its limited handling of emojis. During the testing phase, it became evident that Tool A often displayed high confidence scores, even when participants' personal perceptions differed significantly from the tool's assessment. This lack of contextual sensitivity, combined with its "black box" nature, led to a diminished level of trust in the tools outputs.

Tool B received predominantly positive feedback, mainly due to its simple and transparent explanations alongside sentiment classifications. Participants (P1, P2, P3*, P4*, P6) appreciated that Tool B's analyses often aligned with their own assessments and valued the additional context provided by analyzing the video transcript. However, challenges remained, particularly in detecting humor, and sarcasm. While the detailed reasoning offered by Tool B contributed to its transparency, some participants found the explanations overly lengthy, which occasionally hindered clarity and comprehension.

Tool C stood out for its detailed analysis, offering numerous labels and filter options. While this variety allowed for a more nuanced examination of comments, it also introduced a degree of complexity that some participants (P1, P4*) found overwhelming. Feedback highlighted the confusing arrangement of categories and the occasional ambiguity of certain labels (e.g., “Constructive Criticism”, “Special Flags”, “Cathartic Release”). While some users (P2, P3*, P6*, P7) saw value in the differentiated labeling feature, others (P1, P4*) perceived the tool’s complexity as a barrier and expressed a desire for a clearer, more intuitive interface.

Summary of Findings (User Tasks)

The observations from the user tasks suggest that Tool B provided the most understandable and convincing results, though it requires further fine-tuning to handle linguistic nuances better. In contrast, Tool A was seen as too rigid and lacking contextual awareness, while Tool C, despite its detailed approach, suffered from an overly complex and sometimes unclear presentation.

Additionally, the naming conventions in Tool C’s categories were often unclear, indicating a need for refinement. Interestingly, participants preferred the reasoning provided by the labels over the explanations in the “show analysis” section of Tool C, even though both were identical. This suggests that the way information is presented can significantly impact users’ perceptions of clarity and trustworthiness.

5.4.3 Post-Task Survey: Analysis

After completing the task sheet, participants were asked to complete a survey evaluating the tools in several dimensions, including transparency and trust, effectiveness, complexity and understanding, alignment, and their general impression of the tools. The survey questions can be found in Appendix 7. The next section presents an overview of the survey results, offering insights into participants’ experiences and perceptions.

In the post-task survey, participants were asked to assess transparency and trust across the sentiment analysis tools. Tool B generally received the most positive feedback in this area (see Figure: 5.8). Its detailed explanations, including insights into key phrases, tone interpretation, and comment context, were aimed at fostering trust and were verbally highlighted during the user tasks. Although Tool C provided solid explanations, it did not quite match Tool B’s level of clarity. Tool A, on the other hand, was perceived as less transparent, underscoring the importance of clear communication in building user trust.

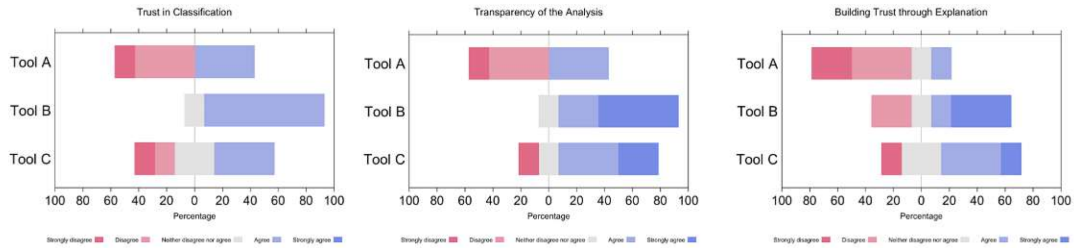

Figure 5.8: **User Evaluation (n=7) of Trust and Transparency in Sentiment Analysis Tools.** This figure shows user evaluations of three sentiment analysis tools, focusing on trust in classification, transparency of the analysis, and building trust through explanations. Tool B and Tool C receive higher trust and transparency ratings, with users expressing stronger agreement regarding the transparency of their analysis and explanatory value. Tool A, by contrast, is rated lower in all three areas.

Effectiveness

Another key dimension assessed in the post-task survey was the effectiveness of the tools, focusing on the comprehensibility and usefulness of the information provided (see Figure: 5.9). Tool B emerged as the most effective, particularly excelling in the clarity of its explanations, which helped users better understand the sentiment classifications and fostered trust in its analysis. While Tool C also performed well, its explanations lacked the same level of clarity, and Tool A, with its reliance on confidence scores alone, was perceived as significantly less effective. During the user tasks, some participants (P1, P2, P6*) even adjusted their interpretation of comments after reviewing the reasoning provided by the tool, demonstrating the impact of clear and detailed explanations on user perception.

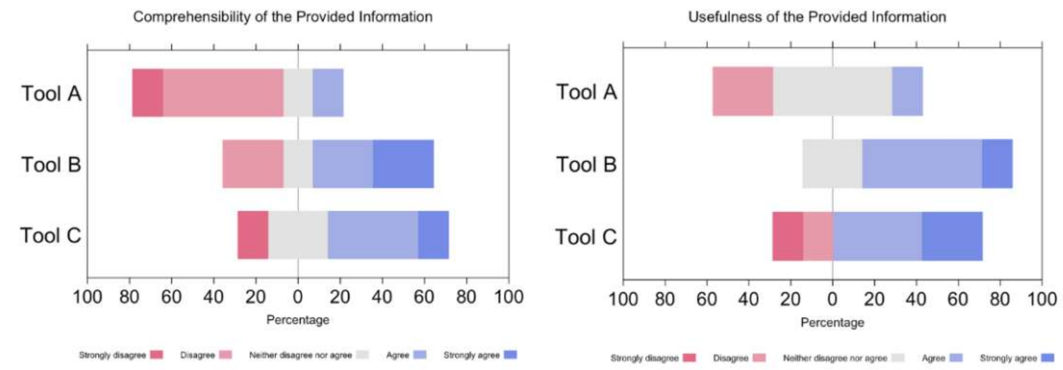

Figure 5.9: **User Evaluation (n=7) of Sentiment Analysis Tools: A Comparison of Tool Comprehensibility and Usefulness of Provided Information.** This figure illustrates user perceptions of three sentiment analysis tools based on their comprehensibility and the usefulness of the information they provide. Tools B and C receive higher ratings for both criteria. Tool A is rated lowest in terms of comprehensibility, with most users remaining neutral regarding its usefulness.

Complexity and Understanding

In the post-task survey, participants were also asked to evaluate the complexity and understanding of the tools (see Figure: 5.10). Tool B emerged as the top performer in this dimension, with participants highlighting its combination of ease of use and clear, easily understandable results. Tool A and Tool C were rated similarly in terms of usability, with both offering a good balance of functionality and simplicity. Overall, none of the tools were considered overly complex, with Tool A and Tool B receiving particularly positive feedback. Although Tool C was perceived as slightly more complex, it still remained within a generally positive range, despite one participant (P4*) expressing frustration during the user tasks. When assessing the usefulness for gaining insights into comment sentiment, Tool B again stood out as the most helpful, followed by Tool C, while Tool A failed to meet participants expectations.

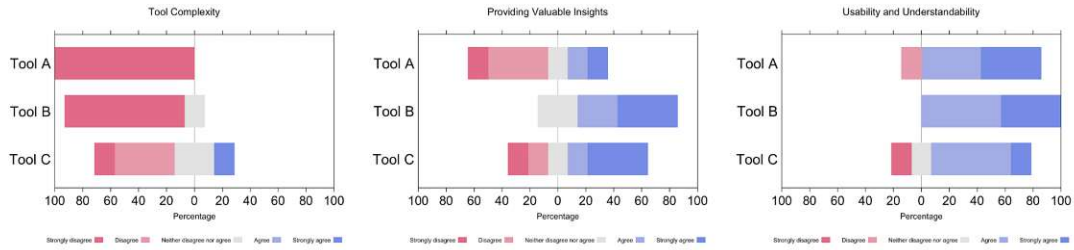

Figure 5.10: **User Evaluation (n=7) of Sentiment Analysis Tools: A Comparison of Tool Complexity, Insightfulness, and Usability.** This figure illustrates user perceptions of three sentiment analysis tools based on their complexity, the value of insights provided, and overall usability and understandability. Tool B receives higher usability and insightfulness ratings. Tool C, though providing valuable insights, is seen as more complex. Tool A, perceived as the least complex, ranks lowest in terms of providing valuable insights.

Alignment

Another key dimension assessed in the post-task survey was alignment (see Figure: 5.11). Alignment refers to how well the tool's sentiment classifications match the participants' own assessments and expectations of the comments. Tool B received the highest ratings, as its classifications closely mirrored what participants expected from the analysis. Participants particularly appreciated Tool Bs ability to understand the nuances of YouTube comments, making its output feel more in line with their own interpretations. Tool C followed in second, although, despite its multi-tier classification system designed to capture more detailed nuances, it was not perceived as fully aligning with participants' expectations. This discrepancy was noted by two participants (P2, P5) during the user tasks, who felt that Tool C's reasoning did not always match the classifications it provided. In contrast, Tool A was less effective in this respect, as all participants found its output to be less aligned with their own perspectives. Overall, these findings suggest that SLMs, like those used in Tool B and Tool C, provide greater sensitivity to contextual subtleties, which aligns better with user expectations compared to the BERT-based model used in Tool A.

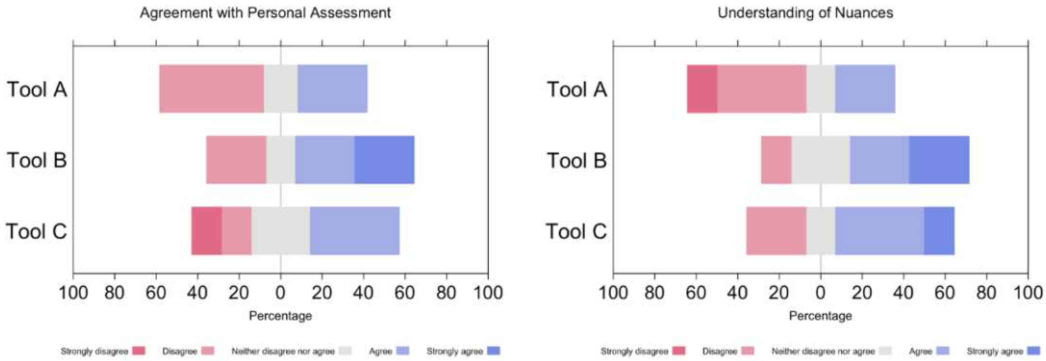

Figure 5.11: **User Evaluation (n=7) of Agreement with Personal Assessment and Understanding of Nuances in Sentiment Analysis Tools.** This figure illustrates user feedback on how well three sentiment analysis tools align with users' personal assessments and handle nuanced interpretations. Tool B and Tool C receive higher agreement ratings, reflecting closer alignment with users own sentiment judgments and better nuance comprehension. Tool A, in contrast, scores lower in both categories, suggesting a perceived lack of nuanced understanding and weaker alignment with users' personal evaluations.

General Impression

In terms of overall satisfaction, Tool B was the clear favorite among participants, receiving more positive feedback compared to Tool A and Tool C (see Figure: 5.12). Tool B resonated well with users and had the highest likelihood of being recommended to others. Tool C received more moderate ratings, showing some potential but not as strongly as Tool B, while Tool A struggled to generate strong support or advocacy.

All three tools contributed to a moderately positive shift in participants' perspectives on the usefulness of automated comment analysis. Tool B again led the way, demonstrating the strongest impact on altering user perceptions, followed by Tool C and then Tool A. These findings highlight Tool Bs effectiveness in delivering valuable insights and building confidence in sentiment analysis technology.

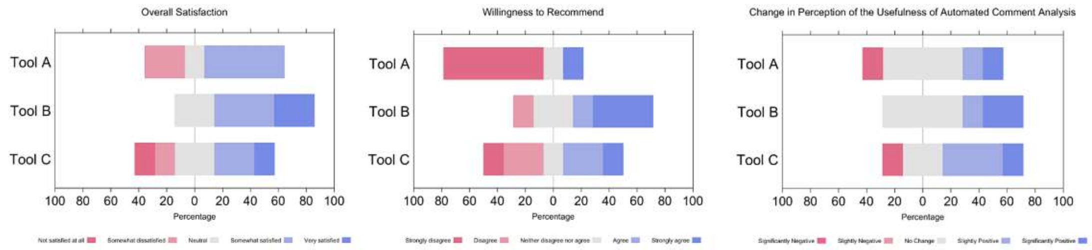

Figure 5.12: **User Evaluation (n=7) of Overall Satisfaction, Willingness to Recommend and Change in Perception of the Usefulness of Automated Comment Analysis in Sentiment Analysis Tools.** This figure illustrates user feedback on the overall satisfaction with three sentiment analysis tools, users' willingness to recommend them, and the change in perception regarding the usefulness of automated comment analysis tools. Tool B received the highest ratings in overall satisfaction, recommendation likelihood, and change in perception. Tool A ranked lowest in willingness to recommend and in perceived usefulness of automated comment analysis but slightly outperformed Tool C in overall satisfaction.

Summary of Findings (Post-Task Survey)

To summarize, Tool B consistently received the highest ratings, particularly for transparency, usefulness, and explanation quality. Its analyses helped users better understand the sentiment classifications of YouTube comments, illustrating how a balance of simplicity, transparency, and interpretability fosters greater user acceptance.

Tool C also showed potential, especially in its ability to interpret nuances, though it was perceived as more complex. In contrast, Tool A, while easy to use, scored poorly in terms of usefulness and transparency.

These results support the hypothesis that SLMs, as used in Tool B and Tool C, offer higher potential for sentiment analysis of YouTube comments compared to the BERT-based approach used in Tool A.

After discussing the analysis of the different parts of the user evaluation study, the next section will present the thematic analysis of the transcripts from the semi-structured interviews conducted and will be complimented with the notes taken and observations made during the user evaluation study.

5.5 Thematic Analysis

The thematic analysis aims to uncover key patterns and insights from participant feedback regarding the sentiment analysis tools tested in the user evaluation study. By systematically analyzing participants' responses as described in Chapter 3, several recurring themes emerged, highlighting both, the strengths and limitations of the tools. The themes identified are **Confidence and Transparency**, **Understanding**, **Information Presentation**, **Cognitive Load**, **Target Audience**, **Feature Requests**, and **Emotion**

Recognition. The following sections discuss each theme in detail, grouped by either one or two high-level categories supported by participant quotes and the assigned codes that form each category.

5.5.1 Confidence and Transparency

Confidence and transparency in sentiment analysis tools seem to play a role in shaping user trust and perceived reliability. All participants questioned overly confident sentiment scores in Tool A which impacted their confidence in the tools' outputs. They where wondering, how the tool (Tool A) arrived at its decisions. This theme consists out of the categories *Overconfidence in Sentiment Scores* and *Lack of Explainability*.

Overconfidence in Sentiment Scores

All participants noted that Tool A, in particular, displayed excessively high confidence scores in sentiment classification (e.g., 99%). This led to skepticism and distrust regarding the accuracy of the results. A frequently raised concern was: "Why is it 99% sure that this is negative?" ("Wieso ist es sich zu 99% sicher, dass das negativ ist?").

Codes: [Unjustified_confidence], [Lack_of_doubt], [Excessive_certainty]

Lack of Explainability

Understanding how a model arrives at a classification influences its perceived trustworthiness [101, 103], as also the results of this study show. Regarding Tool A, all participants stated that they could not comprehend how the tool reached its conclusions—particularly in terms of confidence scores and the resulting sentiment classification. This lack of transparency led to reduced trustworthiness in Tool As classifications, as also illustrated in Figure: 5.8. In contrast, Tool B and Tool C were perceived as more trustworthy, as they both offered a reasoning for their classifications.

Codes: [Unhelpful_confidence], [Not_transparent], [No_understanding]

5.5.2 Understanding

A key factor in user satisfaction was the ability to understand how sentiment classifications were made. All participants valued explanation-based approaches, which not only justified the tools decisions but also helped them refine their own interpretations of the analyzed content. The category which shaped this theme is *Explanation-Based Approaches*.

Explanation-Based Approaches

Providing explanations improved the interpretability of sentiment classifications, helping users understand how and why specific classifications were made. Tool B and Tool C were able to provide explanations for their classification results, which all participants viewed positively. Participants (P1, P3*, P7) remarked: "Oh, Tool B even highlights key phrases that contributed to the classification!" ("Oh, Tool B beschreibt ja sogar Keyphrasen, die zur Klassifikation beigetragen haben"), "Ah, now I understand how the tool interpreted this—yes, that could also be a valid meaning." ("Ah, jetzt verstehe ich, wie das Tool das verstanden hat, stimmt, das könnte natürlich auch so gemeint sein."), and "Yes, thats exactly how I would have interpreted the comment as well." ("Ja, genau so hätte ich den Kommentar auch interpretiert.").

Each explanation consisted of a tone analysis, key phrase identification, and context evaluation, which were combined to justify the models classification decision. Notably, participants (P1, P2, P3*, P5, P6*, P7) actively engaged with these explanations, particularly when they disagreed with a classification. In some cases, this even led them to question their own initial interpretation of a comment (P2, P3, P6*). Furthermore, participants (P1, P2, P4*, P6*, P7) observed that the tool occasionally identified contextual cues from the video that they had overlooked, thereby, according to them, deepening their understanding of the topic.

Codes: [Builds_understanding], [Explanations_preferred], [Key_phrases_helpful], [Questioning_own_understanding], [Precise_justification], [Supports_opinion_formation], [Reasonable_performance], [Deeper_understanding]

5.5.3 Information Presentation

The way sentiment analysis results were presented influenced user engagement and perceived ease-of-use. While some participants (P1, P2, P4*, P5) preferred a simple categorization system, others valued nuanced classifications (P3*, P6*), though excessive detail in Tool C was seen as overwhelming or imprecise (P1, P3*, P4*, P5, P6*). These notions are represented by the two categories *Simple Categorization* and *Value of Nuance*.

Simple Categorization

Participants (P1, P2, P4*, P5) preferred a straightforward classification system that provided high-level sentiment labels such as positive, negative, and neutral (Tool A, B). One participant (P5) stated: "For my purpose, the simple version suffices; I don't need any more details than that." ("Für meinen Zweck reicht das simple, ich brauche gar nicht mehr Details als das."). Further, the ability to access a more detailed explanation (Tool B and Tool C) by clicking on a classification was well received. Participants (P1, P2, P4*, P5) appreciated that the initial display of labels did not show the explanation right away, allowing them to engage with more details only if desired which emphasizes prior work from Shneiderman regarding the *Visual Information-Seeking Mantra* [102].

In contrast, Tool C was criticized for displaying too many labels (P1, P4*, P5). Each comment could have up to four labels, including the main classification result, up to two tone analysis labels, and an optional special flag. This was perceived as visually cluttered and did not resonate with participants (P1, P4*, P5). They expressed that such a level of nuance was unnecessary from their perspective.

Codes: [Nuance_not_needed], [Simple_preferred], [Basic_classification_sufficient]

Value of Nuance

While nuanced classification can provide deeper insights, Tool C received criticism for cluttering the comment section visually. However, all participants acknowledged that it was perceived as a more professional tool, particularly useful for YouTube creators or corporate users. They noted that Tool C offered a deeper analysis, moving beyond a simple classification system, which could be valuable for these specific user groups.

Although one participant (P1) explicitly praised how the explanations show that the tool is able to detect emotions, others (P3*, P4*, P5) argued that the predefined emotion categories (e.g., Neutral/Informative, Humorous/Playful, Sarcastic/Ironic) displayed in the labels were too rigid and could benefit from greater precision. One participant (P4*) asked: “Why are neutral and informative combined? In general, why do they always consist of two words? From my perspective, this doesn’t fit and is too imprecise.” (“Warum ist neutral und informativ zusammen? Generell, wieso bestehen die immer aus zwei Worten? Das passt aus meiner Sicht nicht und ist zu ungenau.”)

Codes: [Benefits_reviews], [Professional_utility], [Emotion_detection], [Higher_precision]

5.5.4 Cognitive Load

During the user evaluation study, cognitive load has shown to be a factor in determining whether participants find a tool intuitive or overwhelming. Participants (P1, P4*, P5) expressed frustration when faced with excessive information, suggesting that a more streamlined presentation of sentiment classifications could enhance ease-of-use. This theme is supported by the category *Information Overload*.

Information Overload

Excessive detail did hinder ease-of-use. In particular, Tool C was perceived as overwhelming due to the way it labeled comments and the extensive filtering options available for each label. Participants (P1, P4*, P5) reported feeling unsure about where to start or how to navigate the tool effectively, despite the fact that its basic interaction patterns remained consistent with Tool A and Tool B. One participant (P4*) explicitly rejected Tool C, stating: “I completely dismiss Tool C. This is too much—way too overwhelming.” (“Tool C lehne ich ab. Das geht gar nicht—viel zu viel.”).

Additionally, a participant (P6*) suggested that explanations, in general, could be more concise. Instead of full-sentence explanations, the participant recommended presenting key information in the form of keywords or bullet points to improve readability and reduce cognitive load.

Codes: [Too_many_details], [Overwhelming_information], [Excess_options], [Overwhelming_nuance], [Shorten_explanations]

5.5.5 Target Audiences

All participants suggested that the usability and impact of the sentiment analysis tools depend on the specific needs of different user groups. Participants (P1, P2, P3*, P5, P6*, P7) identified various potential beneficiaries, including content creators, corporate users, and even parents, while also raising concerns about potential misuse. The category *Beneficial Users* provides more details for that theme.

Beneficial Users

Potential target groups include various stakeholders. Participants mentioned that they could envision YouTube creators (P1, P2, P3, P4*, P5, P7), corporate users (P7), and even parents (P1, P5) as potential beneficiaries of such a tool. Tool C was specifically highlighted as offering the highest level of detail for filtering, making it particularly useful for these groups.

One participant (P2) suggested that such a tool could support individuals with autism by helping them better understand emotional cues in the comment section. However, another participant (P6*) raised concerns about potential misuse—specifically, how Tool C could be exploited by trolls[3] to intentionally filter and target individuals whose comments are classified as *Emotional/Vulnerable*, or related to *Shared Experiences*, *Follow-up Inquiries*, or *Technical Questions*.

Additionally, three participants (P1, P3*, P6*) stated that they would personally use the tool, with a preference for Tool B while mentioning to be willing to explore the capabilities of Tool C (P3*, P6*).

Codes: [Parental_control_application], [Personal_use], [Creator_tool], [Corporate_moderation], [Autism_support], [Trolling_misuse]

5.5.6 Feature Requests

Participants suggested additional functionalities that could enhance the tools' usability and adaptability. While some users proposed features like parental controls (P1, P3*, P5) and search functionality (P6*) for Tool C, some (P1, P2, P3*) also stated that their preferred tool (Tool B) already met their needs. Based on the coding of the transcripts, the category *Desired Features* emerged.

Desired Features

Users expressed interest in various additional functionalities. Interestingly, two participants (P1, P5)—who had not taken part in the user preference study or any of the pilot sessions—suggested a parental control feature for Tool C, allowing users to automatically hide specific comment categories: "Parents could configure this for their children, and

[3]*Troll* is referred to someone who is intentionally provoking or harassesing others online with their interactions in posting or interacting with content.

comments could be automatically hidden." ("Die Eltern könnten das für ihre Kinder voreinstellen und dann werden die Kommentare automatisch ausgeblendet"). In contrast, participants from the user preference study (P3*, P4*, P6*) stated that at no point anything should be hidden.

Another participant (P6*) proposed the addition of a search bar, stating: "It would be cool to have a search function." ("Es wäre cool, wenn man eine Suchleiste hätte")

Additionally, one participant (P4*) expressed a desire to combine features from Tool A and Tool B, suggesting that explanations should be supplemented with a confidence score for greater interpretability.

However, some participants (P1, P2, P3*) found the existing functionalities sufficient. When explicitly asked about desired features, they responded that they liked the tool (in all cases referring to their favorite tool, namely Tool B) as it was and could not think of any further improvements.

Codes: [Add_confidence_scores], [Search_function], [Hide_categories], [No_additional_features]

5.5.7 Emotion Recognition

Accurately detecting emotions such as sarcasm and humor remains a challenge for the evaluated tools. While some participants (P1, P2) were impressed by the tools ability to identify these nuances, others (P3*, P4*, P5, P7) pointed out misinterpretations and limitations, particularly in cases involving emojis. The category *Detection Difficulties* forms this theme.

Detection Difficulties

Detecting sarcasm and humor remains a challenge with mixed success across the evaluated tools. While some participants (P1, P2) found Tool A and Tool B capable of identifying these nuances, others (P3*, P4*, P5) highlighted notable misinterpretations, particularly when emojis were involved.

One participant (P1) expressed being impressed by the perceived ability of Tool A and Tool B to identify sarcasm and humor. However, others (P3*, P4*) were more critical, noting instances where the tools (Tool A, B, C) failed to recognize sarcasm correctly. As one participant (P3*) remarked: "It does not seem to have understood this" ("Das scheint es jetzt nicht verstanden zu haben"). Participants (P1, P2, P3*, P4*, P5) were further highlighting, that Tool A did not take emojis into account.

Codes: [Mixed_success_humor_sarcasm], [Emojis_not_recognized], [Human_interpretation_differs]

5.5.8 Conclusion

The thematic analysis highlights diverse user perspectives, underscoring the importance of explainability, usability, and well-structured information presentation in sentiment analysis tools. While participants valued the in-depth analysis provided by Tool B and Tool C, concerns about information overload (Tool C), lack of transparency (Tool A), and the persistent challenge of detecting sarcasm (Tool A, B, and C) indicate areas for improvement. Additionally, insights into target audiences and feature requests suggest directions for future development.

The user evaluation study highlights key user concerns regarding sentiment analysis tools, emphasizing transparency, user trust, and the trade-off between simplicity and nuance. Future improvements could focus on better explanations and customizable filtering, ensuring a balance between detailed sentiment analysis and ease of use while focusing on one of the proposed target audiences. Overall, these insights contribute to a better understanding of user expectations and provide valuable guidance for refining sentiment classification tools.

CHAPTER 6

Discussion

This chapter reflects on the findings of this master's thesis, connecting the research questions to the findings, methodological considerations, model comparisons, user study insights, ethical implications, research limitations, existing research and future directions.

6.1 Research Questions and Hypotheses

The primary goal of this thesis was to evaluate the performance of Small Language Models (SLMs) compared to a BERT-based model (DistilBERT-sst2) in the task of sentiment analysis of YouTube comments.

Research Question 1: The first research question investigated whether SLMs, despite not being explicitly fine-tuned for sentiment analysis, could achieve better results than DistilBERT-sst2, a BERT-based model specifically fine-tuned for sentiment analysis. The model evaluation results (see Chapter: 4) demonstrated that over 80% of the tested SLMs outperformed DistilBERT-sst2 in sentiment analysis doing binary classification in *zero-context*, 65% in binary classification *with context* and also 65% in three-way classification in *zero-context* and *with context*. Additionally, SLMs aligned more closely with the human baseline in binary (positive/negative) sentiment classification. When extending the task to three-way classification (positive, negative, neutral), the superiority of the SLMs remained, although the macro-F1 scores decreased slightly. Furthermore, the integration of additional context, namely the transcript of the associated YouTube video, revealed that not all models consistently benefited from this additional information. However, some models exhibited a performance boost. This partially supports the hypothesis that most SLMs outperform the chosen BERT-based model without prior fine-tuning, while also highlighting that the "context effect" is model-dependent.

Research Question 2: The second research question focused on user preferences regarding the display of sentiment in YouTube comments. The results indicated that transparency and simplicity are key elements for users. The initial hypothesis—that users would prefer positive comments to be hidden if they were misclassified—was not confirmed. Instead, participants expressed a desire for all content, particularly negative comments, to be displayed rather than hidden.

Research Question 3: This question compared SLMs and the BERT-based model from the perspective of perceived usefulness for users during sentiment analysis. The user evaluation study (see Section: 5.4) confirmed that SLMs are perceived as more useful because they can explain their classification results, even when those results are incorrect. All participants preferred the developed browser extension with the explanatory SLM-based approach over a simple confidence score from the BERT-based model, as these explanations provided a better understanding of the nuances and context of the comments.

6.1.1 Contributions

Previous research presented sentiment analysis results to users on separate webpages [87, 80], through visualizations [23], or by reporting metrics from executed code [2]. This work is novel because it evaluates the performance of locally deployed SLMs for sentiment analysis in comparison to a state-of-the-art fine-tuned BERT-based model on YouTube comments, with a focus on the user as the central recipient of the results. To achieve that, a custom extension for the Google Chrome browser was built, enabling users to perform sentiment analysis while staying on YouTube supporting the field of user-centered content moderation. The following contributions are made:

- **(Technical)** An open-source [1] Chrome extension for YouTube comment moderation, based on insights from user preferences for content filtering.
- **(Technical)** Insights to the feasibility of locally running language models in sentiment analysis are documented.
- **(Human-centered)** An understanding of user preferences and needs in content visualization regarding YouTube comments.
- **(Human-centered)** An understanding of the perceived usefulness when applying a SLM for sentiment analysis instead of a BERT-based model.

6.2 Evaluating the Methodological Approach

The methodological approach employed in this thesis effectively addressed the research questions through a structured, phased process. The research began with a user preference study (see Section: 5.1), utilizing interviews and mock-ups to gather early feedback and inform the design and functionality of the tools. Iterative refinement during pilot testing

[1] On GitHub: `https://github.com/doyouknowmarc/CommentSense` - last accessed 26.03.2025.

led to the development of a cohesive evaluation framework. The user evaluation study (see Section: 5.4), conducted face-to-face, provided insights into user behavior and acceptance of the tools. Concurrently, the model evaluation phase (see Chapter: 4) employed a multi-faceted approach, comparing models against a dataset annotated by participants of the user study in binary and three-way classification, both with and without contextual information, to comprehensively assess model performance.

Despite its effectiveness, the methodology had limitations. Small sample sizes in both user studies, potential biases from differing execution modes (remote vs. in-person), genre-specific dataset constraints, and the limitations of DistilBERT-sst2—such as its binary classification capability, restricted sequence length, and English-only training data excluding emojis—were notable challenges. While these limitations do not invalidate the findings, they highlight areas for improvement in future research.

6.3 Insights from the Model Evaluation and Comparison

The evaluation of sentiment analysis models revealed several key findings. Most SLMs outperformed DistilBERT-sst2 in terms of macro-F1 score, without dedicated fine-tuning for sentiment analysis tasks. Furthermore, all participants in the user evaluation study expressed a preference for SLMs due to their ability to provide explanations for classification results, emphasizing the importance of interpretability in sentiment analysis. Users also responded positively to context-based classification, particularly when models explicitly linked their decisions to the provided context. Although context often improved accuracy and alignment with human judgment, its impact varied between SLMs, underscoring the nuanced role of context in classification performance.

6.3.1 Computational Efficiency vs. Accuracy

The study explored whether a SLM-based approach could outperform a BERT-based model approach, revealing that more than 80% of SLMs achieved better performance in binary *zero-context* and 65% in binary *with context* and three-way *zero-context* and *with context* classification tasks. However, while termed "small", SLMs are computationally larger and less efficient than DistilBERT-sst2. When computational constraints are disregarded, SLMs demonstrate potential for sentiment classification, not only due to their alignment with human-annotated dataset but also due to their ability to be easily adaptable with just a prompt instead of dedicated fine-tuning.

From a computational perspective, DistilBERT-sst2 remains more efficient, but efficiency is of limited value if accuracy is compromised. The evaluation confirmed that SLMs generally provided better classification accuracy, as measured by macro-F1 scores, compared to the human-annotated dataset. However, the impact of additional context varied: some models, like mistral-small:22b, showed improvements, while others did not benefit consistently. Notably, some SLMs underperformed relative to the BERT-based model, and certain outputs were deemed questionable.

The rapid evolution of SLMs and supporting hardware, such as Apples Mac Studio with 512 GB [2] unified memory and NVIDIAs DGX Spark [3] enables the deployment of larger SLMs locally. Based on the evaluation, specific SLMs, such as the Phi-series and Mistral model, are recommended. For instance, Phi-3.5:3.8b delivered strong results with high efficiency, while Phi-4:14b and mistral-small:22b also showed good performance but have a much higher parameter count. Newer models with higher parameter counts generally performed better, though fine-tuning could enable smaller models to outperform larger ones.

6.4 Insights from the User Evaluation Study

The thematic analysis of user feedback from the user evaluation study yielded several valuable insights. Participants expressed a clear preference for a transparent and simple tool, which they perceived as more user-friendly and trustworthy. This finding is supported by the post-task survey results (see Section: 5.4.3). Users emphasized the importance of explainable classification results, as this increased their confidence in the systems outputs. Additionally, different target groups—such as YouTube creators, parents, individuals with autism, and content moderators—were identified as potential beneficiaries of the more complex Tool C. Feedback also revealed that user requirements for comment display varied depending on whether they envisioned the tool for personal use or as a moderation aid for others. However, some participants raised concerns about Tool Cs excessive options and detail, noting that it could increase cognitive load and potentially enable targeted attacks from adversaries. Technical limitations, such as the longer processing times of larger SLMs compared to DistilBERT-sst2, were also noted. To mitigate this, a pre-processing step was implemented to minimize waiting times during the evaluation, ensuring a smoother user experience.

The results from the user evaluation study regarding the sentiment analysis tools (Tool A, B, C) can be summarized as follows:

- Tool A (DistilBERT-sst2) was rejected due to an excessively high confidence score in combination with inaccurate classifications.
- Tool B (SLM-based), with its simple, transparent high-level classification (positive, negative, neutral) and a deeper analysis when needed, provided the best user experience.
- Tool C (SLM-based) was criticized for excessive options and detail while classes where not distinct enough. But, participants acknowledged to perceive Tool C rather as a professional tool than one for personal use.

[2]Needs to be selected in the configuration: https://www.apple.com/at/shop/buy-mac/mac-studio/apple-m3-ultra-mit-28-core-cpu,-60-core-gpu,-16-core-neural-engine-96-gb-arbeitsspeicher-1tb - last accessed 23.03.2025

[3]https://www.nvidia.com/de-de/products/workstations/dgx-spark/ - last accessed 23.03.2025

6.4.1 Trustworthiness of Sentiment Analysis Tools

Transparency in sentiment analysis tools—such as providing clear explanations for classifications and avoiding overly high confident scores—enhances both user trust and perceived reliability. Participants expressed skepticism toward Tool A's lacking transparency, while those offering reasoned outputs (Tool B and Tool C) were viewed as more trustworthy. This preference for transparency was evident in the study, as participants favored Tool B (SLM-based) for its simple, transparent high-level classifications (positive, negative, neutral) and optional deeper analysis, which provided the best user experience. While language models have gained popularity, understanding their decision-making processes in terms of potential hidden objectives [75], alignment faking [49], and how to manipulate features to steer the model effectively [113] remains an open research area, presenting opportunities for future work.

6.4.2 Ethical and Societal Implications

The study highlights several ethical and societal implications of using SLMs for sentiment analysis. A key finding is that the explanatory functions of SLMs can influence users' opinions, underscoring the need to evaluate how models justify their classification decisions to avoid unintended bias or manipulation which is also supported in research regarding models persuasiveness [29]. Transparency in model outputs is essential to ensure users can trust and interpret results accurately.

The proposed browser extension promotes user-centered content moderation by enhancing transparency and empowering users to actively engage with and interpret comments. However, the study reveals a tension between the desire for simplicity in tool design and the need for detailed information to support informed decision-making. Aiming for balance is important for future developments, particularly when tailoring tools to different audiences. Simplified versions may be more suitable for personal use, while feature-rich versions could better serve, e.g., content moderators who require deeper contextual insights.

6.5 Limitations

This thesis has several limitations that should be acknowledged. First, the BERT-based model used in the study has inherent constraints: it is limited to binary classification, processes only English text without emojis, and has a restricted input length for messages. Additionally, the rapid evolution of SLMs during the study period underscores the need for continuous updates to ensure insights into state-of-the-art capabilities.

Second, the dataset and comment selection present limitations. The dataset, shaped by a specific YouTube genre and a limited number of annotated comments, affects the generalizability of the findings. The small sample sizes in the user preference study (n=6) and user evaluation study (n=7) further restrict the applicability of the results to a broader user base.

Finally, differences in the evaluation environment—such as variations between remote and in-person testing—may have introduced inconsistencies in user feedback, complicating the interpretation of results. These limitations highlight areas for improvement in future research.

6.6 Future Work

The findings highlight the importance of transparency and simplicity in sentiment analysis systems, empowering users to manage their comment experiences independently without relying on overbearing platform controls.

An interesting observation was that some users assumed that the classes presented by Tool C during the user evaluation study represented the tools full classification scope, unaware of its broader capabilities. This was likely influenced by the study design, where Tools A and B consistently displayed classes (positive, negative, neutral) they are capable of classifying, leading users to expect the same from Tool C. Additionally, users associated Tool Cs red/orange-ish labels with negativity, mirroring the color coding of Tools A and B, despite differing meanings: blue represented the *primary class*, pink the *tone analysis* and orange *special flags*. This aligns with the spillover effect [91] where the evaluation of the prior tools influenced the evaluation of the latter. Future work could explore randomized testing orders and improved onboarding to better communicate Tool Cs expanded classification scope and minimize such spillover effects.

Building on insights from the user research and the model evaluation, several directions for future research emerge. First, contextual classification needs further exploration: how do classification result alignment change when human annotators also evaluate comments without video context, and what additional inputs—such as likes, user interactions, or creator labels—could further enhance classification accuracy? Second, the design of explanations should be optimized to achieve a balance between providing useful detail and avoiding user overload. Third, adapting the browser extension for different target groups—such as simplified versions for casual users and feature-rich tools for content moderators—could improve usability, acceptance and functionality. Finally, as the field of SLMs evolves, continuous evaluation of emerging SLMs and their capabilities will be essential to maintain state-of-the-art performance.

6.7 Implications

This thesis introduces an innovative approach to moderating YouTube comments by leveraging locally running SLMs capable of explaining their classification decisions. This method offers a more transparent and user-centric solution for content moderation. The approach is transferable to other social platforms, such as Reddit, Twitter, or Twitch, provided the models are regularly updated to reflect state-of-the-art advancements. However, several challenges and open questions remain.

When evaluating models, certain technical issues must be addressed. For instance, the pipeline which is processing YouTube comments into prompts must be robust enough to handle injection attacks, e.g., hidden messages encoded in emojis or special characters which allow malicious actors to manipulate the system [126] but, also needs to be able to process legitimate usage of emojis and special characters. Employing a safety layer to prevent universal jailbreak attacks is also challenging in large language model research [98].

CHAPTER 7

Conclusion

This research explored the potential of on-device Small Language Models (SLMs) in sentiment analysis of YouTube comments, comparing their performance to a state-of-the-art DistilBERT-sst2 model within a Chrome extension. For the user study, three tools were created (Tool A, Tool B, Tool C). The findings demonstrate that SLMs offer distinct advantages, particularly in providing reasoning for classification decisions, which enhanced user trust, transparency and perceived usefulness. The simplified sentiment labeling approach in Tool B was well-received by users, outperforming both the more complex multi-tier classification of Tool C and the confidence score-based approach of Tool A. The model evaluation revealed that SLMs not only align well with human sentiment assessments but also offer flexibility through prompt engineering. This adaptability simplifies the introduction of new classification categories and rules without the need for costly and time-consuming retraining processes, as required by traditional BERT-based models. Additionally, SLMs exhibited a strong capability to interpret emojis, and provide nuanced sentiment reasoning. Further, they offer multilingual capabilities and can be augmented with additional context to enhance their contextual understanding and reasoning. While this research demonstrates promising results, several limitations must be acknowledged. These include potential biases in sentiment classification, the computational constraints of SLMs, and the challenge of balancing accuracy with explainability. Future research could explore refining the models, extending the approach to other social media platforms, or integrating user-defined filtering mechanisms for even greater personalization.

Ultimately, this work contributes to the broader discussion on ethical and user-driven content moderation. By shifting some control back to users and leveraging SLMs, this thesis proposes a step toward more transparent and adaptable sentiment analysis systems in online environments.

Appendix

The appendix of this thesis includes content in both German and English. This is due to the nature of the project, which involved technical documentation, user feedback, and research materials in both languages. To maintain authenticity and accuracy, original German sources, such as survey materials, the task sheet or interview questions have been preserved. English content, particularly in screenshots or when displaying additional data from the evaluation, complements the German material. This bilingual approach reflects the practical context of the study. The codebase of this thesis is open-source and can be found on GitHub: `https://github.com/doyouknowmarc/CommentSense`.

User Preference Study: Interview Guideline

This section presents the questions asked during the semi-structured interview.

Interview Fragebogen

Fragen zum allgemeinen YouTube-Nutzungsverhalten

- Wie oft nutzen Sie YouTube und zu welchem Zweck? (z.B. Unterhaltung, Information, Lernen)
- Welche Art von Videos schauen Sie sich auf YouTube an? (z.B. Musikvideos, Tutorials, Vlogs)
- Lesen Sie die Kommentare unter den Videos? Wenn ja, wie oft?
  - Warum lesen/lesen Sie nicht die Kommentare?
- Haben Sie schon mal einen Kommentar geschrieben? Wenn ja, wie oft?

Fragen zur Interaktion mit Kommentaren

- Welche Arten von Kommentaren finden Sie hilfreich/nützlich?
- Welche Arten von Kommentaren finden Sie störend/unangenehm? (z.B. beleidigende Kommentare, Spam, irrelevante Kommentare)
- Haben Sie schon einmal negative Erfahrungen mit Kommentaren auf YouTube gemacht? Wenn ja, welche?
- Was tun Sie, wenn Sie auf einen negativen Kommentar stoßen? (z.B. ignorieren, melden, selbst einen Kommentar schreiben)
- Würden Sie eine Funktion begrüßen, die negative Kommentare automatisch herausfiltert?
- Welche Kriterien sollten Ihrer Meinung nach für die Filterung von Kommentaren verwendet werden?

Fragen zu den verschiedenen Anzeigemethoden von Kommentaren

- Welche der folgenden Anzeigemethoden für Kommentare finden Sie am hilfreichsten/übersichtlichsten?

  (Prototypen in einem Screenshare zeigen)

  - Gruppierung von Kommentaren nach Sentiment (positiv, neutral, negativ)

  - Hervorhebung von Kommentaren mit negativem / positiven Sentiment
  - Ausblenden von Kommentaren mit negativem Sentiment
  - Andere Anzeigemethoden (bitte erläutern)
- Welche Auswirkungen hätte die automatische Filterung von Kommentaren auf Ihr YouTube-Erlebnis? (z.B. positiver, negativer, keine Auswirkung)

Fragen zur Browser-Erweiterung "CommentSense"

- Welchen Browser benutzen Sie um YouTube Videos zu schauen?
- Wären Sie bereit, eine Browser-Erweiterung zu nutzen, die Ihnen hilft, negative Kommentare auf YouTube zu filtern?
- Welche Funktionen sollte eine solche Browser-Erweiterung Ihrer Meinung nach haben & warum?
- Welche Bedenken hätten Sie im Hinblick auf die Nutzung einer solchen Browser-Erweiterung? (z.B. Datenschutz, Genauigkeit der Filterung)
- Was würden Sie bevorzugen:
  - ein positiver Kommentar der als negativ interpretiert wird und deswegen nicht angezeigt wird / gehighlighted wird
  - ein negativer Kommentar der als positiv interpretiert wird und deswegen angezeigt wird / gehighlighted wird
- Wie Transparent soll die Erweiterung für Sie sein (nachvollziehbarkeit der Ergebnisse anzeigen welche Inhalte ausschlaggebend waren, erläutern)

Fragen zur Ethik

- Welche ethischen Bedenken haben Sie bei der Verwendung eines solchen Tools?
- Wie würden Sie auf die Bewertung einer Ihrer Kommentare reagieren?
  - Ein positiver Kommentar von Ihnen wird nicht angezeigt, weil er als Negativ erkannt wurde

User Evaluation Study: Surveys, Task Sheet, Interview Guideline, Study Setup

This section presents the questions used in the pre-task and post-task surveys. It also includes the task sheet and the questions asked during the semi-structured interview. Finally, it describes the lab setup during the study.

Pre-Task Survey Questions

1. Bitte geben Sie Ihr Alter an
2. Höchster Bildungsabschluss
3. Bitte wählen Sie das Niveau, das Ihre Englischkenntnisse am besten beschreibt.

Optional Bitte geben Sie Ihr Geschlecht an.

4. Wie oft nutzen Sie YouTube?
5. Zu welchem Zweck nutzen Sie YouTube hauptsächlich? (Mehrfachauswahl möglich)
6. Lesen Sie die Kommentare unter YouTube-Videos?
7. Warum lesen Sie die Kommentare? (Mehrfachauswahl möglich)
8. Warum lesen Sie selten bis nie Kommentare? (Mehrfachauswahl möglich)
9. Wie hilfreich fänden Sie ein Tool, das YouTube-Kommentare automatisch analysiert und filtert?
10. Welche Bedenken hätten Sie grundsätzlich gegenüber der Nutzung eines solchen automatisierten Tools?

Aufgabenblatt

Bitte denken Sie daran, Ihre Gedankengänge bei der Ausführung der einzelnen Aufgaben laut auszusprechen.

Aufgabe 1 - Vorbereitung

YouTube Training Video [1]

YouTube Video: Want a Winning Smile? Learn How to Cheat at Cards! [2]

Sie werden ein 2:22 YouTube Video sehen, in dem Jason Ladanye einen Kartentrick vorführt, welcher von einem Zuschauer angefragt wurde. Seine Aufgabe ist es, zwei Karten (Königin und Ass) aus dem Kartendeck zu ziehen und dann auf magische Art noch einen König erscheinen zu lassen.

a) [ ] Bitte schauen Sie das Video bis zum Schluss. Geben Sie ein kurzes Signal, wenn das Video zu Ende ist.

b) [ ] Bitte lesen Sie die eingeblendeten Kommentare (ohne Kommentar-Antworten) aus der Kommentarsektion (die ersten 20).

c) [ ] Kommentieren Sie das Gelesene bitte kurz und weisen Sie auf etwaige Kommentare hin, die Ihnen aufgefallen sind.

Aufgabe 2 - Tool A

a) [ ] Wechseln Sie in den nächsten Tab (TOOL A). Dort ist das gleiche YouTube Video geöffnet. Gehen Sie direkt in die Kommentarsektion.

b) [ ] Verschaffen Sie sich einen Überblick über die Analyse-Ergebnisse des Tools und werfen Sie auch einen Blick auf den Konfidenzscore. Weisen Sie bitte auf etwaige Punkte hin, die Ihnen auffallen.

c) [ ] Nutzen Sie bitte die Filterfunktion (positiv, negativ, neutral). Erläutern Sie Ihre Gedanken dazu kurz.

Aufgabe 3 - Tool B

a) [ ] Wechseln Sie in den nächsten Tab (TOOL B). Dort ist das gleiche YouTube Video geöffnet. Gehen Sie direkt in die Kommentarsektion.

[1] Introducing the Tools and how to use them: https://www.youtube.com/watch?v=FQxMjcuW1-M - last accessed 23.03.2025

[2] Study video: Want a Winning Smile? Learn How to Cheat at Cards! https://www.youtube.com/watch?v=rnCjM_sovp0 - last accessed 23.03.2025

b) [ ] Verschaffen Sie sich einen Überblick über die Analyse-Ergebnisse des Tools und werfen Sie auch einen Blick in die Begründungen. Weisen Sie bitte auf etwaige Punkte hin, die Ihnen auffallen.

c) [ ] Nutzen Sie bitte die Filterfunktion (positiv, negativ, neutral). Erläutern Sie Ihre Gedanken dazu kurz.

Aufgabe 4 - Tool C

a) [ ] Wechseln Sie in den nächsten Tab (TOOL C). Dort ist das gleiche YouTube Video geöffnet. Gehen Sie direkt in die Kommentarsektion.

b) [ ] Verschaffen Sie sich einen Überblick über die Analyse-Ergebnisse des Tools und werfen Sie auch einen Blick in die Begründungen. Weisen Sie bitte auf etwaige Punkte hin, die Ihnen auffallen.

c) [ ] Nutzen Sie bitte die Filterfunktion (By Sentiment, By Tone, By Special Flag). Erläutern Sie Ihre Gedanken dazu kurz.

Post-Task Survey Questions

1. “Das Tool war einfach zu bedienen und die Ergebnisse waren für mich gut nachvollziehbar.”
2. “Die bereitgestellten Informationen haben mir geholfen, die Stimmung der Kommentare richtig einzuschätzen.”
3. “Die Darstellung der Ergebnisse (z. B. Begründung oder Konfidenzscores) ermöglichte mir ein klares Verständnis der Analyse.”
4. “Ich empfand das Tool als unnötig komplex, sodass es schwierig war, die Ergebnisse zu interpretieren.”
5. “Bewerten Sie Ihre Gesamtzufriedenheit mit dem Tool”
6. “Bewerten Sie Ihre Gesamtzufriedenheit mit dem Tool” [Kommentar]
7. “Das Tool versteht die Nuancen von YouTube-Kommentaren.”
8. “Ich vertraue darauf, dass das Tool YouTube-Kommentare korrekt klassifiziert.”
9. “Das Tool liefert nützliche Einsichten in die Stimmung der YouTube-Kommentare.”
10. “Ich würde dieses Tool für die Analyse von YouTube-Kommentaren weiterempfehlen.”
11. “Die Klassifikationen des Tools entsprechen meiner eigenen Einschätzung der Kommentare.”
12. “Die bereitgestellte Begründung (Reasoning oder Konfidenzscore) hat mir gehoflen, die Entscheidungslogik des Tools zu verstehen.”
13. “Ich fand die Begründung des Tools überzeugend.”
14. “Die Begründung hat mein Vertrauen in die Analysen des Tools erhöht.”
15. “Hat die Nutzung des Tools Ihre ursprüngliche Einschätzung hinsichtlich der Nützlichkeit einer automatisierten Kommentaranalyse verändert?”

Semi-strukturiertes Interview

Durchführung nachdem die Testperson die Tools getestet hat und den Fragebogen ausgefüllt hat.
Dauer: ca. 15-20 Minuten.

Basierend auf folgenden Hypothesen:

- Hypothese 1: SLMs liefern bessere Sentimentanalysen
- Hypothese 2: Reasoning statt reiner Konfidenzscores
- Hypothese 3: Nuancierte Darstellung wird einfacher Klassifikation bevorzugt

Für Hypothese 1

a) "Wie bewerten Sie die Qualität der Sentimentanalysen, die Sie während des Tests erlebt haben?"

b) "Welche Unterschiede haben Sie zwischen den Tools festgestellt, insbesondere im Hinblick auf die Genauigkeit der Begründung (Reasoning / Konfidenzscore)?"

Für Hypothese 2

a) "Inwiefern hat Ihnen die Begründung (Reasoning) dabei geholfen, die Analyseergebnisse zu verstehen?"

b) "Würden Sie ein Tool bevorzugen, das explizit seine Entscheidungslogik erklärt, anstatt lediglich Konfidenzscores anzuzeigen? -> Warum oder warum nicht?"

Für Hypothese 3

a) "Wie wichtig ist Ihnen eine differenzierte bzw. nuancierte Darstellung der Kommentare im Vergleich zu einer einfachen Positiv/Negativ/Neutral-Klassifikation?"

b) "Können Sie Beispiele nennen, bei denen Ihnen eine nuancierte Analyse besonders weitergeholfen hat oder weiterhelfen würde?"

Vergleichsfragen

a) "Welches der getesteten Tools hat Ihnen insgesamt am besten gefallen und warum?"

b) "Gibt es Funktionen oder Darstellungsweisen, die Sie sich zusätzlich gewünscht hätten?"

Abschließende Fragen

a) "Würden Sie das Tool nutzen?"

b) "Für wen denken Sie, wäre dieses Tool hilfreich?"

Environment and Setup

The following section displays the setup during the in-person study at TU Wien. It shows the position of the participant (yellow) and of the researcher (purple). Further it highlights how the devices and documents where set up.

Figure 1: The study environment setting during the user evaluation study. During user tasks, the keyboard was removed from the participant. In the interview setting, the seating changed.

Miscellaneous

This section presents the developed tools used to support model evaluation and transcribe the interviews. Additionally, it includes the creator's approval to use their face and name in the screenshots. It also showcases the user interface of Tools A, B, and C. Finally, it displays the classified data from the IMDb dataset.

Streamlit Dataset Evaluation Tool

The following section displays the dataset evaluation tool which was used for evaluating and comparing sentiment results of for example the IMDb dataset.

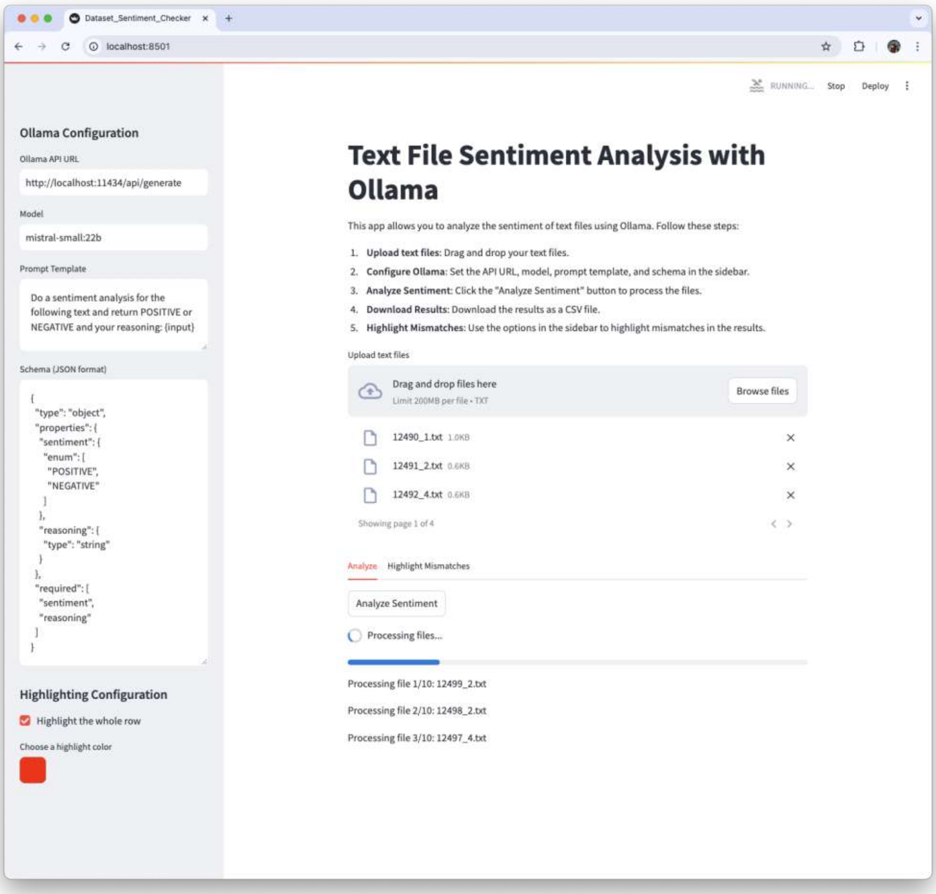

Figure 2: The custom-built web interface used for evaluating pre-classified datasets against classification results from SLMs during the thesis. The tool is publicly available here: "https://huggingface.co/spaces/doyouknowmarc/dataset-evaluation-helper"

Transcription Tool

The following section displays the transcription tool which was used to transcribe the interviews.

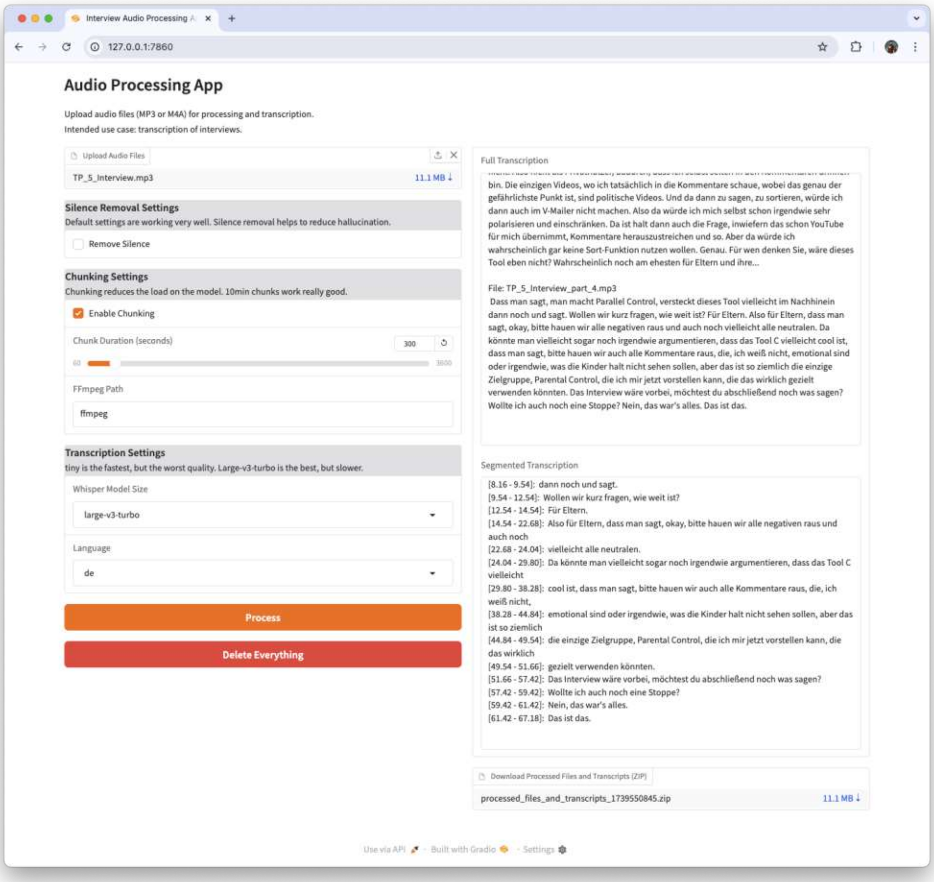

Figure 3: The custom-built web interface used for segmenting and transcribing the interview transcripts during the thesis. The tool is publicly available here: "https://huggingface.co/spaces/doyouknowmarc/Transcription"

Approval to use Jason Ladanyes face and channel name

This section displays the approval to use the name and the face of the YouTube creator *Jason Ladanye*.

Bobbi
Aw: Contact form
An: Marc Kroll
Eingan...@gmail.com 1. März 2025 um 18:10

Hi Marc,

Thanks for reaching out!

Yes, you can use the video as is without blurring Jason's face or channel name.

Thanks
Bobbi

From: Website Enquiry <jason@cardmagicbyjason.com>
Date: February 25, 2025 at 9:15:21 AM EST
To: Jason Ladanye <jason@cardmagicbyjason.com>
Subject: Contact form

From: Marc Kroll @gmail.com
Phone Number:
Subject: Should I blur your face and name in my master thesis?

Message:
Hi,
first of all: I'm a fan.

Second:
During my master thesis I used the following video:
https://youtu.be/rnCjM_sovp0?feature=shared&t=1
during an user evaluation study.

Users had to use a set of tools I developed (Chrome Extensions) to analyze the sentiment of your comment section.

To illustrate the tools, screenshots were taken, which not only show the tool, but also you.
To be precise, the exact frame you can see when following the link above (second 1 of the video "Want a winning smile..")

Of course, usernames and userpictures from the comment section are blurred.

-> Should I do the same for your name and face?

Kind regards,
from Vienna, Austria,
Marc

Figure 4: Creator approval

User Interfaces: Tool A, B, C

The following section displays the user interface of Tool A, Tool B and Tool C.

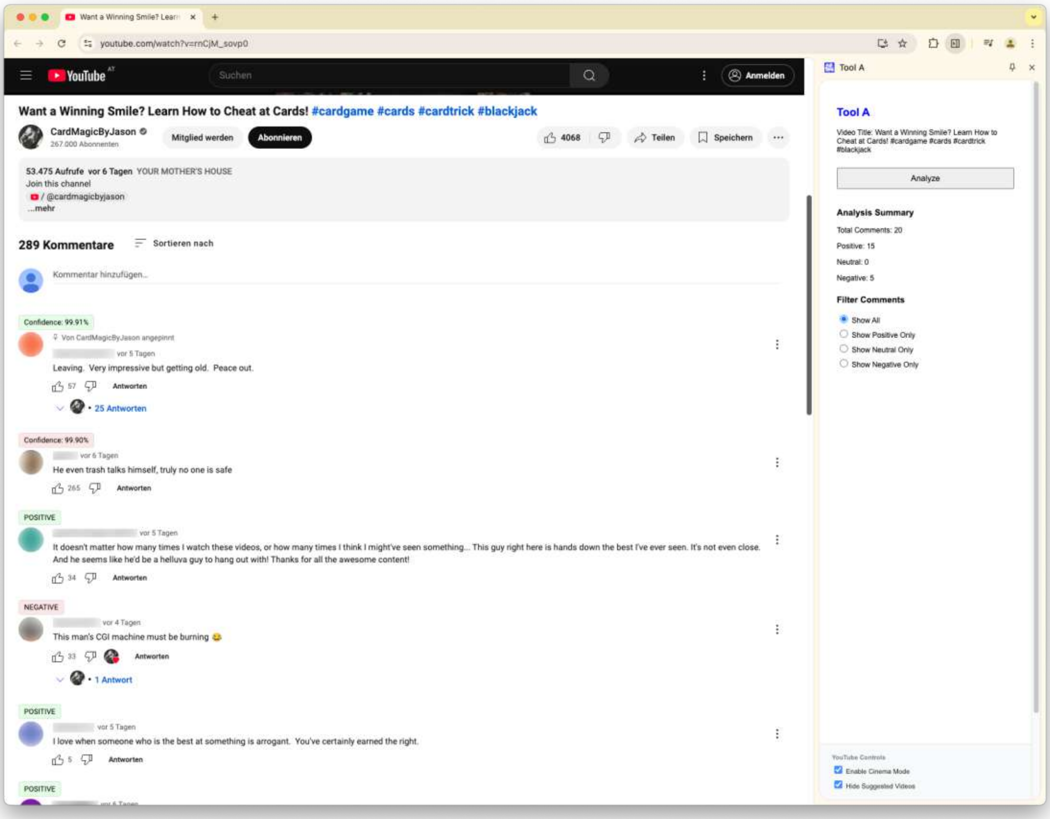

Figure 5: Tool A: Sentiment analysis using DistilBERT-sst2. It performs binary classification (positive, negative) and comments are tagged and a sentiment score is provided.

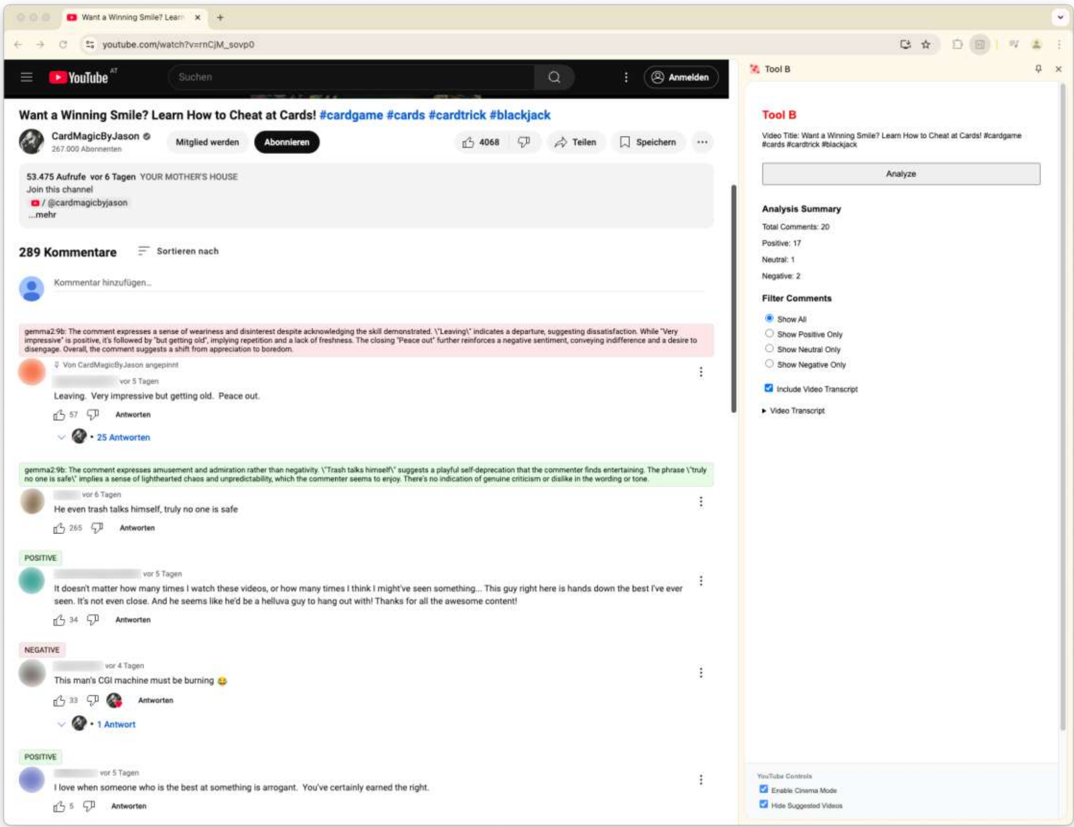

Figure 6: Tool B: Sentiment analysis using a SLM. It performs three-way classification (positive, negative, neutral) and provides explanations for its classifications.

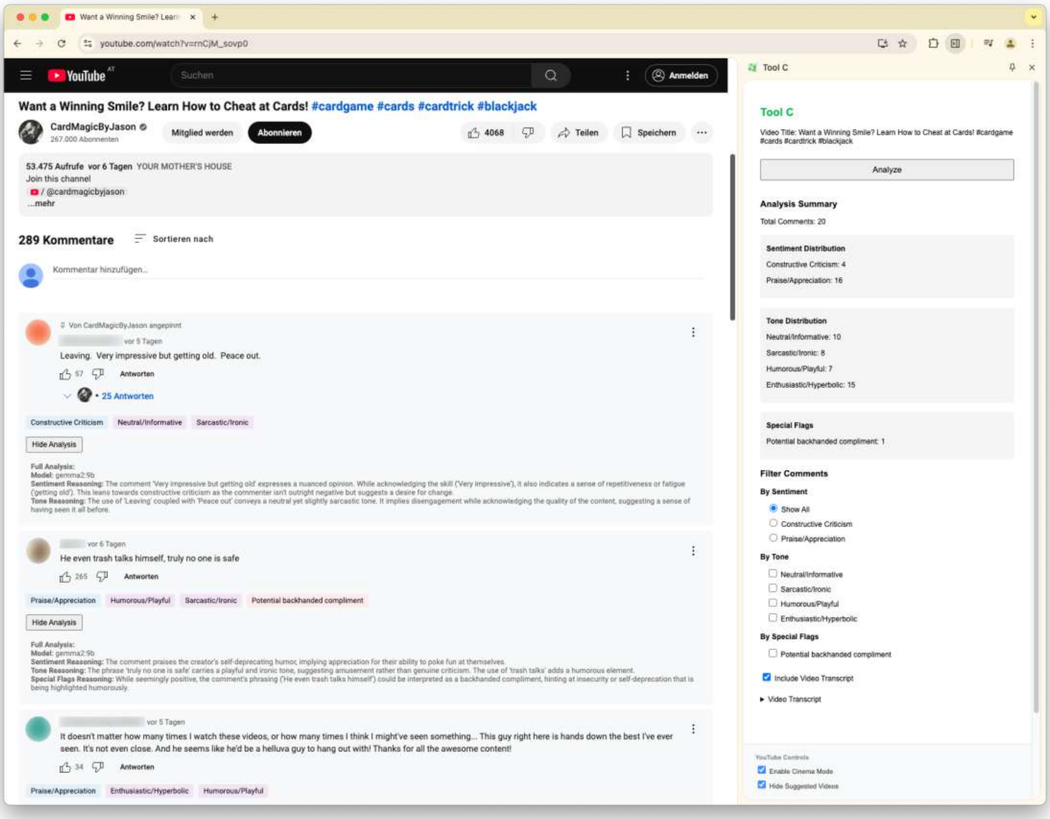

Figure 7: Tool C: Performs multi-tier classification and provides it's explanation.

Multi-tier Classification Classes: Tool C

Below is the multi-tier classification system listed. It first determines the primary class (sentiment) then the secondary class (tone) and afterwards potential special flags. For all classes, the model is instructed to provide an explicit explanation, why it has chosen that class.

- **sentiment** (string, required):
  - Praise/Appreciation
  - Constructive Criticism
  - Video Quality Feedback
  - Content Correction
  - Personal Compliment
  - Creator Advice/Suggestion
  - Collaboration Request
  - Debate/Argumentation
  - Shared Experience
  - Inside Joke/Community Reference
  - Technical Question
  - Follow-Up Inquiry
  - Educational Addition
  - Nostalgia/Throwback Reaction
  - Cathartic Release
  - Outrage/Moral Judgment
  - Spam/Promotion
  - Off-Topic
  - Cultural Reference
  - Meta Commentary
  - Misinformation Flag
- **tone** (array, required, 1-2 items):
  - Enthusiastic/Hyperbolic
  - Sarcastic/Ironic
  - Neutral/Informative
  - Emotional/Vulnerable

  - Humorous/Playful
  - Critical/Harsh
  - Nostalgic/Sentimental
  - Controversial/Provocative
  - Aggressive/Hostile
  - Sympathetic/Supportive
  - Confused/Disoriented
  - Ambiguous/Uncertain
  - Formal/Professional
  - Casual/Informal

- **special_flags** (array):
  - Potential backhanded compliment
  - Cultural/language ambiguity
  - Virtue signaling
  - Attention-seeking behavior
  - Multi-layered meaning
  - Hate Speech
  - Harmful
  - Emoji-Driven
  - Political/Religious Reference

- **reasoning** (object, required):
  - **sentiment** (string, required): Explicit justification for sentiment classification
  - **tone** (string, required): Explanation of tone choices with examples
  - **special_flags** (string, required): Explicit justification of any flagged elements

The initial classification schema was created using DeepSeek-V3 on January 7th 2025 with the query listed in Listing 1 which resulted in Listing 2 and represented the first draft for a multi-tier classification system which was then build upon.

Listing 1: Query to DeepSeek V3 to enhance the existing prompt

```
please help me to improve the following prompt:

    prompt = f"""
    Your goal is to analyze a comment from the YouTube comment
        section and to classify it into one class.
```

```
    Decide for a class that best fits the comment. You have to
       decide for one class.
    Explain your reasoning.

    Here is the comment:
    "{comment}"
    """
```

Listing 2: Response from DeepSeek V3 for enhancing the prompt

```
Here's an improved version of your prompt with clearer structure,
   better guidance, and enhanced specificity:

prompt = f"""
Analyze and classify the following YouTube comment by following
   these steps:

1. **Classification Task**:
   Categorize the comment into ONE primary category from these
      options:
   - Praise/Appreciation
   - Criticism/Disagreement
   - Question/Inquiry
   - Suggestion/Feedback
   - Personal Experience
   - Humor/Sarcasm
   - Spam/Promotion
   - Off-Topic
   - Other (specify reason)

2. **Analysis Requirements**:
   - Identify key phrases, tone, and context
   - Consider both literal meaning and implied intent
   - Acknowledge if multiple categories could apply, but select
      the MOST DOMINANT one

3. **Response Format**:
   **Classification**: [Selected Category]
   **Reasoning**:
   - Bullet-point analysis of linguistic cues and contextual
      evidence
   - Explanation of why this category was chosen over similar
      alternatives

4. **Special Cases**:
   - Flag harmful/hateful content immediately
   - Note ambiguous language or cultural references that might
      affect interpretation
```

```
Example of a good response:
**Classification**: Humor/Sarcasm
**Reasoning**:
- Uses hyperbolic emoji combination
- References popular meme format ("This video cured my fear of
   ...")
- Lacks substantive feedback expected in other categories

Now analyze this comment:
"{comment}"
"""

Key improvements made:
1. Added explicit categories to reduce ambiguity
2. Structured the analysis process with clear steps
3. Specified output format for consistency
4. Included examples to guide response quality
5. Added handling of edge cases (harmful content, cultural
   references)
6. Emphasized requirement to compare/contrast similar categories
7. Added linguistic analysis requirements (tone, phrasing, intent)

This version will produce more consistent, explainable, and
   nuanced classifications while maintaining focus on the core
   task.
```

Transcript

The following section displays the transcript from the YouTube video [3].

```
all right this one got my attention for
$500 the request the request how about
you learn how toing talk you all right
this one got my attention for $500 this
request is for Dr Elana a
top cosmetic
dentist in San Diego feels like I'm
advertising pull a queen and an ace
together from the
waterfall oh that's easy and then uh
have the King of Hearts Leap out of the
deck well we're going to have to fire up
the old CGI machines over there you guys
hear that King from the deck um pulling
a queen and an ace together that's
pretty
easy uh apparently doing running cuts to
the table is difficult because we just
drop some cards that's all right
then um let's make it
happen Ace and a queen together Shuff is
legit
obviously uh nothing on top nothing on
the bottom no Ace Queen nothing in my
hands um yeah you good Ace and queen let
me just touch my face and palm of Ace
and queen all right that's
easy now this next
bit uh getting a king of hearts the one
and only King of Hearts to LEAP out of
the deck it's going to be difficult
because generally cards don't leap out
of a deck but we'll try this I
can't I wasn't really paying attention
to how many times these cards were
shuffled but we'll try something before
we continue only two
```

[3] Study video: Want a Winning Smile? Learn How to Cheat at Cards! https://www.youtube.com/watch?v=rnCjM_sovp0

```
cards I will push these cards into the
deck like
this and I will push down like this and
one card will
leap from the top of the deck and that
is in
fact the king of
I am so good anyway that's 500
bucks
$500 and uh if you're in San Diego and
you want a top cosmetic dentist there's
only one place to
go it's that easy
folks it's that that easy welcome to the
hustle
```

IMDb Dataset

The following section displays the classifications and reasonings of the small language model Phi4:14b for 20 randomly selected entries of the *negative* IMDb training dataset in Table 1. Notably, the model is classifying some of the comments as positive or neutral instead of negative. Future work could explore all of the 25000 entries from the IMDb training dataset to evaluate its correctness.

Filename	Comment	Sentiment	Reasoning
12435_3.txt	Yes AWA wrestling how can anyone forget about this unreal show. First they had a very short interviewer named Marty O'Neil who made "Rock n Roll" Buck Zumhofe look like a nose tackle. Then it was Gene Okerland who when he got "mad as the wrestler" would say either "Were out of time" or "Well be right back" acting like he was mad but actually sounding forced. After he went to the WWF Ken Resneck took over even though his mustache looked like week old soup got stuck to it was a very fine interviewer who "Georgeous" Jimmy Garvin called mouse face which made me fall off my chair laughing. After he jumped ship then Larry Nelson came on board which he was so bad that Phyllis George would of been an improvement! Then there's Doug McLeod the best wrestling announcer ever who made every match exciting with his description of blows! Then he was offered more pay by the Minnesota North Stars hockey team. At ringside who can forget Roger Kent who's mispronouncing of words and sentences were historic Like when a wrestler was big "Hes a big-on!" punched or kicked in the guts "right in the gussets"or when kicked "He punted him" or "the "piledriver should be banned" after Nick Bockwinkle used it on a helpless opponent.(Right Roger like you care!) After he left to greener money(WWF) they had Rod Trongard who's announcing style was great but different. Like when a wrestler scraped the sole of his boot across another guys forehead he'd say"Right across the front-e-lobe" or when a wrestler is in trouble "Hes in a bad bad way". He also would say AWA the baddest,toughest,meanest, most scientific wrestlers are here right in the AWA!(No extra money Verne Gagne!) After he left(WWF) Larry(Wheres Phyllis?!) Nelson took over and I would talk to someone else or totally ignore him.(WWE wisely didn't take him!) Also Greg Gagne had the ugliest wrestling boots I ever saw a yellow color of something I don't want to say.Also when hes looking for the tag he looks like he wants to get it over with so that he can run to the nearest restroom! Jumpin Jim Brunzell was such a great dropkick artist that you wonder why Greg was ever his partner. Jerry Blackwell(RIP)was also a superstar wrestler but you wonder why Verne had himself win against him.(Puhleeeeze!) Then when Vince McMahon would hire Gagnes jobbers, he would make most of them wrestle squash matches. I like to see the Gagne family say wrestlings real now!	POSITIVE	The text overall reflects a fond and nostalgic sentiment towards AWA wrestling despite its humorous criticism of certain announcers and wrestlers. The language used is playful, with comedic elements highlighting memorable moments or quirks (e.g., Marty O'Neil making Buck Zumhofe look like a nose tackle, Ken Resneck's mustache humor). Positive remarks are made about Doug McLeod being the best wrestling announcer ever and Rod Trongard having a great announcing style. The text also includes appreciation for certain aspects of AWA wrestling, suggesting that it was an 'unreal show' with historical significance in the wrestling community. Despite some negative comments on specific individuals or events, the overall tone suggests enjoyment and affection for the past era of AWA.

12434_4.txt	I saw this movie, just now, not when it was released and hailed as best picture of the year here in Israel. and to summarize everything right now, I will just say: this is not a good film. This is Dror Shaul's second feature film, and I have to admit that his first and the TV drama he made before this picture are much better. further more, this is his first attempt at directing a drama. the early works were comedies, and were funny and effective. The first thing you have to know if you'll ever see this film: Israel of the 21st century hates the kibbutz and the values it represented since the formation of the state of Israel. the real situation of the kibbutzim is very dire, and some of them disappear one by one. the kibbutz, Hebrew word for collective, was a sort of village for members only, where the values of equality and socialism were the dogma for everyday life. with the change in social values with time, it seems now that the kibbutz was a place where the human spirit was repressed, locked within the dogma rules, with no ticket out. the entrance of capitalist values and way of life in the 90's and so far made it very hard on the kibbutzim to survive. the crazy mother in the film is the central metaphor for that. But, I regard this film as having nothing to do with nostalgia for the good old days of the kibbutz. once, it was a dream of every young couple to live in a kibbutz and raise children in this quite and beautiful environment. but the film shows the opposite. that the kibbutz, with it's socialist dogma, was a place sort of like a cult of crazy people, with crazy ideas that undermine the freedom of each individual within the collective. this is the central philosophy of post modern capitalism: your individuality is the most important thing. you must place yourself in the center, and no one else but you is the matter. this is the philosophy the film stands for, and that's just it's first sin. If you disagree with me on the political side, I'm sure you will agree that the acting, the tone of the film, it's script and it's direction are the four sins that follow. the film has no real visual text and none of it's shots is something to remember. it is also very "delicate", a delicacy that is no more than artsy fartsy attempt to provoke emotions, which do not surface, not in the film and not with the viewer. it brings nothing but boredom. Can someone please explain: why this film won so many prizes? maybe because it shows that Israel is in line with the rest of the world, hating socialist and human values? or maybe it shows that Israel is a "delicate" place, not giving in to dogmas and fanaticism? that we are basically very human and good people, capable of emotions, especially when they are fake ones, just like capitalism expects us to be? or maybe because it tells one of the biggest lies of Israeli cinema in recent years, a lie that undermines the justification of the existence of the Jewish state? no matter what the answer is, it's not a good one. not for the world, not for human values and not for the Jews.	NEUTRAL	The text expresses mixed feelings about the film. The initial statement is somewhat negative ('can't promise much'), criticizing the stereotypical and inaccurate depiction of the kibbutz, which suggests disappointment. However, it also acknowledges positive aspects, such as good acting and beautiful views of Israel. Overall, the criticism seems balanced with appreciation for certain elements, resulting in a neutral sentiment.

12433_3.txt	I saw this movie, just now, not when it was released and hailed as best picture of the year here in Israel. and to summarize everything right now, I will just say: this is not a good film. This is Dror Shaul's second feature film, and I have to admit that his first and the TV drama he made before this picture are much better. further more, this is his first attempt at directing a drama. the early works were comedies, and were funny and effective. The first thing you have to know if you'll ever see this film: Israel of the 21st century hates the kibbutz and the values it represented since the formation of the state of Israel. the real situation of the kibbutzim is very dire, and some of them disappear one by one. the kibbutz, Hebrew word for collective, was a sort of village for members only, where the values of equality and socialism were the dogma for everyday life. with the change in social values with time, it seems now that the kibbutz was a place where the human spirit was repressed, locked within the dogma rules, with no ticket out. the entrance of capitalist values and way of life in the 90's and so far made it very hard on the kibbutzim to survive. the crazy mother in the film is the central metaphor for that. But, I regard this film as having nothing to do with nostalgia for the good old days of the kibbutz. once, it was a dream of every young couple to live in a kibbutz and raise children in this quite and beautiful environment. but the film shows the opposite. that the kibbutz, with it's socialist dogma, was a place sort of like a cult of crazy people, with crazy ideas that undermine the freedom of each individual within the collective. this is the central philosophy of post modern capitalism: your individuality is the most important thing. you must place yourself in the center, and no one else but you is the matter. this is the philosophy the film stands for, and that's just it's first sin. If you disagree with me on the political side, I'm sure you will agree that the acting, the tone of the film, it's script and it's direction are the four sins that follow. the film has no real visual text and none of it's shots is something to remember. it is also very "delicate", a delicacy that is no more than artsy fartsy attempt to provoke emotions, which do not surface, not in the film and not with the viewer. it brings nothing but boredom. Can someone please explain: why this film won so many prizes? maybe because it shows that Israel is in line with the rest of the world, hating socialist and human values? or maybe it shows that Israel is a "delicate" place, not giving in to dogmas and fanaticism? that we are basically very human and good people, capable of emotions, especially when they are fake ones, just like capitalism expects us to be? or maybe because it tells one of the biggest lies of Israeli cinema in recent years, a lie that undermines the justification of the existence of the Jewish state? no matter what the answer is, it's not a good one. not for the world, not for human values and not for the Jews.	NEGATIVE	The text expresses strong negative opinions about the film in question. The author describes it as 'not a good film' and criticizes various aspects such as its acting, tone, script, direction, and visual style. They also express dissatisfaction with its thematic approach to the portrayal of kibbutzim and criticize what they see as an endorsement of postmodern capitalist philosophy. Furthermore, the writer questions the awards it received, suggesting reasons that align with a critical view on the film's cultural significance. Overall, the language is heavily negative throughout.

12432_1.txt	Tainted look at kibbutz life This film is less a cultural story about a boy's life in a kibbutz, but the deliberate demonization of kibbutz life in general. In the first two minutes of the movie, the milk man in charge of the cows rapes one of his calves. And it's all downhill from there in terms of the characters representing typical "kibbutznikim." Besides the two main characters, a clinically depressed woman and her young son, every one else in the kibbutz is a gross caricature of well	NEGATIVE	The text provides a highly critical view of the film's portrayal of kibbutz life, describing it as deliberately negative and filled with harmful stereotypes. The writer expresses disappointment and frustration with how characters are depicted in exaggeratedly villainous terms and describes various elements of the storysuch as cult-like practices and hypocrisyas negative portrayals. Additionally, the reviewer criticizes the film's overdramatization and insinuates that it unfairly demonizes kibbutz life, implying a biased or sensationalized perspective.
12431_1.txt	There just isn't enough here. There a few funny spots, but not really enough. I was very disappointed because I love stupid movies. I was expecting this to be a hidden gem. IT WAS NOT. It was a hidden turd. The whole time I was watching it I was thinking, "Geez, I could do better than that and I'm a moron". Don't believe the few good reviews. I was suckered in by them. At no point did I laugh at loud. Maybe this was funny or cutting edge in 1987 but man, it fell flat today. This movie had promise, it just didn't deliver on it. Maybe if they rewrote it a few more times. I cant fault the actors. They weren't the best, but they just didn't have much with which to work. This movie failed because of threadbare writing. SORELY disappointed.	NEGATIVE	The text expresses significant disappointment and dissatisfaction with the movie. Key phrases indicating negative sentiment include 'There just isn't enough here,' 'I was very disappointed,' 'IT WAS NOT,' 'suckered in by them,' 'it fell flat today,' 'failed because of threadbare writing,' and 'SORELY disappointed.' The author mentions a lack of laughter, unmet expectations, and criticism of the movie's writing quality. Despite acknowledging that the actors weren't bad, the overall tone remains negative due to these criticisms.
12430_1.txt	Mark Pirro's "Deathrow Gameshow" of 1987 is a black comedy that is extremely cheesy in many parts, but occasionally very funny nevertheless. This movie could certainly have been a lot better, the acting is terrible, and some extremely cheesy scenes make it hard to watch at times, but the concept is funny, and it has some hilarious moments. In the near future (the year 1991), game shows have changed. Chuck Todean (John Mc Cafferty) hosts a game show called "Live Or Die", in which convicted death row inmates have the chance to play for their lives, and for money. Candidates who fail, get executed on the air using many different methods, such as guillotines, electric chairs, and other, more bizarre devices of execution, followed by applause from the cheering studio audience. The show is, of course, more than controversial, and Chuck has made lots of enemies... "Deathrow Gameshow" is incredibly cheesy and crappy in many aspects, and the acting is terrible, but it is without doubt fun in many parts, especially if you're a fan of dark humor. You haven't missed anything if you haven't seen it, but it is definitely funny and a good time waster. 4/10	NEUTRAL	The text presents a balanced view by highlighting both the negative and positive aspects of 'Deathrow Gameshow.' On one hand, it criticizes the movie for being cheesy, having terrible acting, and containing scenes that are hard to watch. These elements contribute to a negative sentiment. On the other hand, the text acknowledges that the concept is funny and there are hilarious moments, suggesting some enjoyment, which adds positive sentiment. The overall tone suggests mixed feelings, leaning towards neutral as it recognizes both flaws and entertainment value.

12429_1.txt	I rented this film to see what might be a bloody, non stop action movie and got this overly sentimental and super cheap low budget action-drama that makes Kickboxer look like Die Hard. Lou and Reb are in Vietnam and as Lou saves Reb from the gooks, he gets shot in the head in what is easily one of the worst effects ever. The Vietnam scenes are shot in someones backyard, I swear! Lou is now brain damaged and Reb and him live together and own a bar. Super homoerotic. Lou is convinced to fight in a cage for money and Reb goes on a killing spree to get him back. There is no good fight scenes at all, the punches are two inches away from a person. Characters personalities change in matter of seconds. One guy is a bad and in the next scene he's good. The acting is horrid and the music is some overly sentimental Frank Stallone sounding song that would make you sick. I hated this film.	NEGATIVE	The text expresses strong dissatisfaction with various aspects of the film, including its action scenes, special effects, production quality, character development, acting, and music. The use of words such as 'overly sentimental,' 'super cheap low budget,' 'worst effects ever,' 'horrid,' and phrases like 'makes Kickboxer look like Die Hard' indicate a significant negative impression. Additionally, the statement 'I hated this film' directly conveys intense dislike.
12428_2.txt	Cage (1989) was another one of those low budget "buddy" action flicks that were produced during the 80's thanks in large part due to the films such as 48hrs. and Lethal Weapon. This one stars Reb Brown and Lou Ferrigno as to former Vietnam Vets who happen to run a local dive bar. Reb takes care of Lou because he saved his life in 'Nam. But Lou was shot in the head and is now pretty soft. Although he's huge, Lou has the brain of a child. One day some ruffians throw their wait around in the bar and Lou and Rebb beat the tar out of them. But payback's a mother. They crash the bar leaving Lou and Reb with nothing. That is until these two thugs come into the picture (one of them's a real nice guy) who have a plan in mind. The film's a waste of time. Maybe if they went all they way and made a hard core action flick instead of trying to tone down the gruesomeness of the situation perhaps it could have worked. Alas, it doesn't and the audience is left holding the bag. Oh well. It's too bad because you have all the elements for a great B-movie. Better luck next time, I guess. Not recommended. xxx	NEGATIVE	The text expresses dissatisfaction with the film 'Cage (1989)'. The reviewer criticizes it as a 'waste of time' and notes that it failed to deliver on its potential by not being sufficiently hard-core, suggesting disappointment with how the elements were executed. Phrases like 'maybe if they went all the way', 'alas, it doesn't', and 'oh well' convey a sense of missed opportunity and unmet expectations. The conclusion 'not recommended' further solidifies the negative sentiment.
12427_1.txt	If you want to laugh like crazy, rent Cage. Cage is about two war heroes, Billy and Scott who are best friends. When Billy is shot in Vietnam, he is unable to fend for himself, so Scott takes him in. I have never seen a movie with more gay references to the two main characters. Billy and Scott love to "wrestle" and Scott tells Billy that he is "still sore from last night," among other things. Wonderful catch phrases like "Shut the sh!t up" and "Ping Pang Pong, cut the sh!t" will keep you laughing for hours. The native American guys that are supposed to be playing Mexican gang members are also top notch. As they say, it's "party time right now. Ba-ba-ba-ba ba-ba." I could go on forever, but just watch this movie and laugh your a$$ off. It was so funny I went out and bought the DVD for $5.99	POSITIVE	The text expresses strong enthusiasm and enjoyment regarding the movie 'Cage.' The author uses phrases like 'laugh like crazy,' 'never seen a movie with more gay references' (indicating humor rather than offense), 'keep you laughing for hours,' 'it's so funny,' and 'laugh your a$$ off.' Additionally, they mention purchasing the DVD at an affordable price ($5.99) as a testament to their enjoyment. These elements contribute to an overall positive sentiment.

12426_1.txt	I don't know what you guys are inhaling, but please stop. This movie was complete trash. Macaulay Culkin as Kevin McCallister was lovable, cute, sweet, loving, and extremely fun to watch. He had that innocence but at the same time, he was strong as well. This new kid, Alex? Annoying as all hell! Just a blatant ripoff at trying to match Culkin's charm and grace, failing miserably at that. The villains, complete crap. Harry and Marv, they were such an entertaining on screen team. They were comedic, they were humorous, and just made you really want to root for Culkin even more, which is what a good villain is suppose to do. These guys Joe Pesci and Daniel Stern, absolutely PERFECT for those roles. On the other hand, in 3, we have 4 new idiots that just bicker and again, try to capitalize on the charisma and duo charm of the first villains. Not saying they were horrible, but they were not memorable or even really that entertaining. More on the corny side actually. Then we got the actual movies. Part 1 and 2, classics. Memorable. Perfect to watch around Christmas and brings a wonderful Christmas feeling and brings out so many emotions. It makes you fall in love with the story, and care about the characters. Simply, memorable and close to being masterpieces. Part 3? Masterpiece of crap. Enough said. All and all, I don't know why people like this. I think people really don't know what a good movie is anymore. Its a crappy sequel rip-off with more unrealistic traps and a stupid new main character kid that I wish the robbers would of at least beat the crap out of. Run, run away from this garbage. And lets not even start with Part 4!	NEGATIVE	The text predominantly expresses dissatisfaction and disappointment with the movie being discussed. Phrases like 'complete trash,' 'annoying as all hell,' 'blatant ripoff,' 'complete crap,' and 'masterpiece of crap' highlight strong negative sentiment. The author criticizes nearly every aspect, including character development (new kid Alex) and villains in the sequel compared to the originals. Even when acknowledging positives like Macaulay Culkin's performance or aspects of the first two movies, these are overshadowed by harsh critiques of the newer sequels. Additionally, the concluding advice to 'run away from this garbage' reinforces a negative stance.
12425_2.txt	Whoever thought that ANOTHER Home Alone film would be a good idea should have their head examined... Same plot, different kid, more villains (which leads to MORE endless stupidity in the traps). The other two films were bad enough, and this is where it hits rock bottom. People may as well watch the other films for plot, as it's all identical.	NEGATIVE	The text expresses strong dissatisfaction with the new 'Home Alone' film. It uses phrases like 'should have their head examined,' indicating disapproval of the decision to create another installment in the series. The mention of a 'Same plot, different kid, more villains' and 'MORE endless stupidity in the traps' suggests criticism of repetitiveness and lack of originality. Furthermore, describing this film as hitting 'rock bottom' underscores the negative sentiment by implying it is the worst in the series. Lastly, suggesting that people should watch other films for the plot because they are identical reinforces a sense of disappointment with the new addition.

12424_1.txt	I hated this movie. It was absolutely horrible, poor,poor, PITIFUL acting, REAL REAL REAL stupid criminals that weren't even the LEAST BIT funny(unlike the first 2 home alone movies that were very good). all the boobie traps are weak, pathetic excuses for ideas poorly copied of of the first two which just meant that the people writing this movie were just lazy because their paycheck didn't go above 20 bucks a week. This movie is absolutely lousy, it's not worth even renting. In fact don't even watch it on t.v.! Go use your eyes in a more useful way by seeing the first two! I BARELY give this 1 star(out of 10). Just trust me when I say, if you liked the first two, and you are not a complete stupid person, you will not like this movie,do not watch this movie!!!	NEGATIVE	The text conveys strong negative emotions towards the movie. The author uses harsh language such as 'hated,' 'absolutely horrible,' and 'pitiful acting' to describe their experience. The criticism extends to various elements of the film, including character development ('REAL REAL REAL stupid criminals'), plot devices ('weak, pathetic excuses for ideas'), and effort from the writers ('lazy'). Additionally, the author explicitly advises against watching or renting the movie, suggesting alternatives instead, which further underscores a strongly negative sentiment.
12423_2.txt	I loved the first two movies, but this movie was just a waste of time and money (for me and the studio). I'm still wondering why they made this horrible movie. The thing with the plastic gun and with the toy car, that can go into another house are ridiculous. Joe Pesci and Daniel Stern in the first two movies were so funny, but the terrorists in this one are so stupid and not funny. Believe me this movie is just a waste of time.	NEGATIVE	The text expresses clear dissatisfaction with the third movie in the series compared to the first two. The author describes it as a 'waste of time and money' for both themselves and the studio, indicating disappointment. They express confusion about why such a movie was made and label specific elements like the plastic gun and toy car scenes as 'ridiculous.' Additionally, they contrast the humor in the earlier movies with unamusing characters in this one, referring to them as 'stupid' and not funny. Overall, these expressions convey strong negative feelings towards the film.
12422_1.txt	Remembering the dirty particulars of this insidiously vapid "movie" is akin to digging into your chest cavity with a rusty, salted spoon. Perhaps "Home Alone 2: Lost in New York" (1992) was a bit on the predictable side, but this pathetic excuse for a film is just one of the most shameless bids at commercialization I have ever heard of. A boy fighting off spies/terrorists when he's home alone in a Chicago suburb with the chickenpox? Ridiculous! Why did this film have to be made? I am the kind of person who believes even terrible movies are not wastes of time, but rather learning experiences. However, this is actually a waste of time. It should be avoided at all costs.	NEGATIVE	The text contains several strong negative expressions and criticisms regarding the film 'Home Alone 2: Lost in New York.' Phrases such as 'akin to digging into your chest cavity with a rusty, salted spoon,' 'pathetic excuse for a film,' 'most shameless bids at commercialization,' and 'ridiculous' indicate a highly unfavorable view. The writer emphasizes the negative sentiment by stating that the film should be avoided 'at all costs' and labels it as a 'waste of time.' Even though the text briefly mentions a belief in learning from terrible movies, this does not mitigate the overwhelmingly critical tone.

12421_4.txt	Home Alone 3 is one of my least favourite movies. It's the cream of the crop, or s*** if you tend to be more cynical, as it ranks up (or down) there with stuff like Battlefield Earth and Flinstones: Viva Rock Vegas. In fact, it could even be worse than those two, since those two at least intermittently made me laugh at their stupidity. This just made me cringe in pain constantly and clap when the credits started rolling. No other movie has made me cringe in pain. Now I will point out exactly why this movie is so incredibly atrocious. First off, the plot is ridiculous. It revolves around a chip in a remote control car (?!) that is misplaced and how these terrorists want it. Dumb stuff. The action that ensues is similar to that of the other two Home Alones, with boobytraps and all, but watching these boobytraps being executed is, rather than being funny, incredibly unpleasant to watch. I didn't laugh (or even so much as smile) once, rather, I cringed constantly and hoped that the terrorists would nail the kid. The bird, rather than providing comic relief, was unfunny and annoying. The acting, as done by a bunch of no names, ranges from poor to atrocious. There is not a single good performance here. Alex D.Linz is absolutely unlikeable and unfunny as the kid, while the terrorists act (and judging by their movie credits, look) as they've been hastily picked off the street...and well, that's it. I can see some people saying: "Man, it's for the kids. Don't dis it, man." Well MAN, kids may like this, but they can get a hell of a lot better. See Monsters Inc. and Toy Story before even considering getting this out. Hell, even Scooby Doo and Garfield (which suck - see those reviews for more) are better than this! So in short, this is an irredeemably atrocious movie. This was clearly recycled for the money, as it almost completely rips off the first two; the only thing is, it completely insults the first two as well. No human, kid or otherwise, should find any reason to see Home Alone 3. Ever. It's THAT bad. 0/5 stars	NEGATIVE	The text expresses strong negative opinions about 'Home Alone 3.' The author describes it as one of their least favorite movies and calls it atrocious multiple times. They criticize various aspects of the film, including its plot, action sequences, acting quality, and overall entertainment value. Phrases like 'creams of the crop' (sarcastically) or 'sh***,' 'cringe in pain constantly,' 'unfunny and annoying bird,' 'poor to atrocious acting,' and explicit statements such as 'No human, kid or otherwise, should find any reason to see Home Alone 3. Ever. It's THAT bad.' all contribute to the overwhelmingly negative sentiment.

12420_3.txt	I would like to say that curiosity got the best of me. If only I saw a trailer, I'd be able to tell you the whole plot of the movie; I could have saved myself the most pointless one hour and forty minutes in my entire life, and about twenty dollars. This movie was a disaster waiting to happen, and it is an embarrassment to Hollywood. The movie displays a vivid ignorance of reality. For example, this kid's remote control race car goes all over the neighborhood, and even enters this house. It's even covered with clothes. Is it not rational to believe that a remote can no longer transmit a signal under those circumstances? Hollywood obviously did not believe so. Common logic and any concept of electronics dictates the opposite; I doubt the race car could even have reached the street, let alone a house across the street. Another unrealistic trait is the lack of intelligence the criminals possess. Why is it in all these movies, these criminals are rocket scientists until they encounter an eight-year old? The kid is meant to be the most intelligent person in the movie instead of professional terrorists? Please, there's more reality in The Matrix. Also, the leader puts down his real pistol and "accidentally" picks up a plastic pistol; apparently, he could not tell the difference. Even with a glove on, one should be able to do that. Just because they look the same does not mean anything; there is other senses then sight. The traps are unrealistic as well; if any one of them actually worked, the criminals would be dead. But, Hollywood intends for us to "laugh" at the "funny" results of the traps. I did not laugh; I sighed and rolled my eyes. But, I recommend this movie to anyone who thinks they have seen a really bad movie; the movie they saw will seem like Citizen Kane compared to this one. Otherwise, skip this one for the sake of your pride. Home Alone 3 also raises a question. We all know Alex D. Linz stars in the movie, but did he write it as well?	NEGATIVE	The text expresses strong dissatisfaction with the movie, using phrases like 'most pointless one hour and forty minutes in my entire life' and describing it as a 'disaster waiting to happen' and an 'embarrassment to Hollywood.' The author criticizes various aspects of the film, including its lack of realism and intelligence portrayed by characters. They also mention that they did not find humor or enjoyment in the movie's elements meant to be funny. Additionally, there is an implication that watching the movie could negatively impact one's pride.
12419_1.txt	This is the worst ripoff of Home Alone movies that I have EVER seen! Watch part 1 and two, but don't let anyone say that this is BETTER than the first two! I mean, really, you don't make a movie, then make a sequel with the same characters and actors, and then make another sequel with DIFFERENT characters and actors! I mean, it would have been OK if this wan't a "Home Alone" movie, but they DID make it a Home Alone movie. Culkin is too old now, so you're suppose to STOP making sequels! Goodness, this movie makes me SICK! Buy part 1 and 2.	NEGATIVE	The text expresses strong dissatisfaction with the movie in question. The author describes it as 'the worst ripoff of Home Alone movies,' indicating a very negative opinion. Words like 'ripoff,' 'worst,' and 'makes me SICK' further emphasize the displeasure. Additionally, there is criticism about changing characters and actors in sequels, which suggests disappointment with these decisions. The recommendation to buy only parts 1 and 2 rather than this movie indicates a preference for earlier entries over the current one.

12418_4.txt	I like to keep my reviews short and simple, but this pretty much sums it up. You can not beat the original two for a number of reasons one of which including the directing talent of Chris Columbous. This movie had terrible directing covered up by even more terrible acting. I cant even believe these people are considered actors. Painful to sit through and watch. The storyline was a complete joke about a secret chip and Russian terrorists on a painstaking quest to get it back. Horrible, rent one of the original tow and enjoy yourself! The movie wasn't even set during Christmas like the original. Home Alone was turned from an excellent Christmas time family comedy movie to a joke with no moral or plot!	NEGATIVE	The text expresses dissatisfaction and criticism of the movie. The reviewer finds the directing 'terrible,' the acting 'even more terrible,' and describes the storyline as 'a complete joke' about a nonsensical premise involving Russian terrorists and a secret chip. Furthermore, they mention that it lacks moral or plot compared to the original, which they consider an excellent Christmas family comedy. The overall tone is one of disappointment, frustration, and disapproval.
12417_1.txt	In 1987, John Hughes wrote and directed 'Planes, Trains and Automobiles', which was a hilarious and poignant comedy	NEGATIVE	The text expresses a strong negative sentiment towards John Hughes' work on the sequel to 'Home Alone.' The language used includes criticism such as describing the film as 'uninspired and sadistic,' indicating disapproval of its content. There is also a sarcastic tone when mentioning the introduction of female characters among the crooks, implying that it's not seen as genuinely innovative or positive. Moreover, the review criticizes the lack of charm in the new kid compared to Macaulay Culkin and suggests Hughes' repeated use of tired routines without improvement. Overall, the language conveys dissatisfaction and disappointment with Hughes' work on this particular sequel.
12416_3.txt	Alex D. Linz replaces Macaulay Culkin as the central figure in the third movie in the Home Alone empire. Four industrial spies acquire a missile guidance system computer chip and smuggle it through an airport inside a remote controlled toy car. Because of baggage confusion, grouchy Mrs. Hess (Marian Seldes) gets the car. She gives it to her neighbor, Alex (Linz), just before the spies turn up. The spies rent a house in order to burglarize each house in the neighborhood until they locate the car. Home alone with the chicken pox, Alex calls 911 each time he spots a theft in progress, but the spies always manage to elude the police while Alex is accused of making prank calls. The spies finally turn their attentions toward Alex, unaware that he has rigged devices to cleverly booby-trap his entire house. Home Alone 3 wasn't horrible, but probably shouldn't have been made, you can't just replace Macauley Culkin, Joe Pesci, or Daniel Stern. Home Alone 3 had some funny parts, but I don't like when characters are changed in a movie series, view at own risk.	NEUTRAL	The text provides both positive and negative perspectives on the film 'Home Alone 3'. On one hand, it acknowledges that the movie had some funny parts, indicating a positive aspect. However, it also expresses dissatisfaction with replacing Macaulay Culkin as well as other key figures from previous films, suggesting a negative sentiment towards these changes. Additionally, there is an overall lukewarm evaluation in the statement 'Home Alone 3 wasn't horrible, but probably shouldn't have been made,' which further supports a neutral sentiment by balancing mild criticism with acknowledgment of its passable quality.

Table 1: Three-way Classification done with Phi4 on a supervised trainingset of IMDb negative comments.

Overview of Generative AI Tools Used

The entire thesis was written using Overleaf [4], an online collaborative LaTeX editor. Within this environment, I occasionally used the integrated Writefull [5] tool for grammar checks and minor text refinements. However, Writefull's suggestions were often not useful and, therefore, sparingly applied. I strictly avoided modifying any citation text.

Research Exploration and Conceptual Development

During the initial proposal phase, I used multiple AI assistants to support research ideation and conceptual understanding:

- ChatGPT [6]
- DeepSeek [7]
- Claude [8]
- NotebookLM [9]
- Google AI Studio [10]

These tools were used for:

- Brainstorming potential research directions
- Identifying potential research gaps
- Clarifying complex terminology and concepts

[4] https://de.overleaf.com/ - last accessed 24.03.2025
[5] https://www.overleaf.com/learn/how-to/Writefull_integration - last accessed 24.03.2025
[6] https://chatgpt.com/ - last accessed 24.03.2025
[7] https://www.deepseek.com/ - last accessed 24.03.2025
[8] https://claude.ai/new - last accessed 24.03.2025
[9] https://notebooklm.google.com/ - last accessed 24.03.2025
[10] https://aistudio.google.com/prompts/new_chat - last accessed 24.03.2025

Writing and Language Support

Throughout the thesis writing process, ChatGPT, DeepSeek and Claude assisted in:

- Sentence concision
- Translating from German to English
- Stylistic refinements for academic writing
- LaTex support for formatting
- Restructuring of chapters for better readability

Technical Writing and Development

For code development and technical documentation, I leveraged AI assistants including Claude, ChatGPT, DeepSeek and Gemini [11] to:

- Generate boilerplate code for Chrome extension development
- Establish FastAPI server configurations
- Develop JavaScript for web scraping and DOM manipulation
- Debug technical implementations
- Generate code documentation

Supplementary Technological Tools

Apple's built-in speech-to-text functionality was used to dictate most of the thesis, with subsequent manual or AI assisted refinement. OpenAI's Whisper [12] *large-v2* was employed for local interview transcription.

Example Workflows

Exploration and Ideation

NotebookLM supported me during research by allowing me to upload papers, generate concise summaries, and produce audio overviews for efficient comprehension. It helped to quickly grasp key insights by reading through the summaries and listening to the generated audio while taking notes. Additionally, the platform enabled me to ask clarifying questions, which enhanced my understanding of the material.

[11] `https://deepmind.google/technologies/gemini/` - last accessed 24.03.2025
[12] `https://openai.com/index/whisper/` - last accessed 24.03.2025

Writing

I used Apple's speech-to-text to dictate sections in German or English, guided by my notes. The transcriptions were manually reviewed, then translated and refined for academic writing using AI tools like ChatGPT, Claude, DeepSeek, and Gemini. Except when explicitly stated these tools were not used to generate content that did not originate from me.

List of Figures

List of Tables

Bibliography

[1] Agbesi, V. K., Chen, W., Yussif, S. B., et al. Pre-trained transformer-based models for text classification using low-resourced ewe language. *Systems 12*, 1 (2024), 1. DOI: https://doi.org/10.3390/systems12010001.

[2] Ahuja, H., Kaur, N., Kumar, P., and Hafiz, A. Machine Learning based Sentiment Analysis of YouTube Video Comments. In *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)* (Oct. 2023), pp. 1–6.

[3] Alexa Carboni, B., Sabrina Woessner, B., Olnita Martini, M., and Marroquin, N. A. Natural weight loss or "ozempic face": demystifying a social media phenomenon. *LETTERS TO THE EDITOR* (2024), 1367. DOI: https://doi.org/10.36849/JDD.7613.

[4] Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., et al. Smollm2: When smol goes big – data-centric training of a small language model, 2025. https://arxiv.org/abs/2502.02737 [Accessed: 23.03.2025].

[5] America's Got Talent. Early Release: Madilyn Bailey Sings a Song Made of Hate Comments - America's Got Talent 2021, July 2021. https://www.youtube.com/watch?v=0-V278v9F9s [Accessed: 29.10.2024].

[6] Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. Power to the people: The role of humans in interactive machine learning. *AI magazine 35*, 4 (2014), 105–120. DOI: https://doi.org/10.1609/aimag.v35i4.2513.

[7] Arslan, Y., Allix, K., Veiber, L., Lothritz, C., Bissyandé, T. F., Klein, J., and Goujon, A. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion proceedings of the web conference 2021* (2021), pp. 260–268. DOI: https://doi.org/10.1145/3442442.3451375.

[8] Ashida, M., and Komachi, M. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (2022), pp. 11–23. DOI: https://doi.org/10.18653/v1/2022.woah-1.2.

[9] Badri, N., Kboubi, F., and Habacha Chaibi, A. Abusive and hate speech classification in arabic text using pre-trained language models and data augmentation. *ACM transactions on Asian and low-resource language information processing 23*, 11 (2024), 1–28. DOI: https://doi.org/10.1145/3679049.

[10] Bai, J., Nie, J.-Y., and Paradis, F. Using language models for text classification. In *Proceedings of the Asia Information Retrieval Symposium, Beijing, China* (2004), vol. 42, Citeseer.

[11] Bozdag, E. Bias in algorithmic filtering and personalization. *Ethics and information technology 15* (2013), 209–227. DOI: https://doi.org/10.1007/s10676-013-9321-6.

[12] Braun, V., and Clarke, V. Using thematic analysis in psychology. *Qualitative research in psychology 3*, 2 (2006), 77–101.

[13] Braun, V., and Clarke, V. Conceptual and design thinking for thematic analysis. *Qualitative psychology 9*, 1 (2022), 3. DOI: https://doi.org/10.1037/qup0000196.

[14] Brooks, J., Reed, D. M., and Savage, B. Taking off with a pilot: The importance of testing research instruments. In *ECRM2016-Proceedings of the 15th European Conference on Research Methodology for Business Management": ECRM2016. Academic Conferences and publishing limited* (2016), pp. 51–59.

[15] Catalanello, R., and Sanders, K. Meta is ending its third-party fact-checking partnership with us partners. heres how that program works. *Poynter* (2025). https://www.poynter.org/fact-checking/2025/meta-ends-fact-checking-community-notes-facebook/ [Accessed: 23.03.2025].

[16] Ceci, L. Biggest social media platforms by users as of april 2024. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ [Accessed: 20.10.2024].

[17] Ceci, L. Distribution of removed youtube comments worldwide as of 2nd quarter 2024, by removal reason. https://www.statista.com/statistics/1133165/share-removed-youtube-video-comments-worldwide-by-reason/ [Accessed: 21.10.2024].

[18] Ceci, L. Number of youtube video comments removed worldwide from 3rd quarter 2018 to 2nd quarter 2024. https://www.statista.com/statistics/1132989/number-removed-youtube-video-comments-worldwide/ [Accessed: 20.10.2024].

[19] Chandrasekharan, E., Gandhi, C., Mustelier, M. W., and Gilbert, E. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction 3*, CSCW (2019), 1–30. DOI: https://doi.org/10.1145/3359276.

[20] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al. Scaling Instruction-Finetuned Language Models, Dec. 2022. http://arxiv.org/abs/2210.11416 [Accessed: 30.10.2024].

[21] Chung, T. S., Wedel, M., and Rust, R. T. Adaptive personalization using social networks. *Journal of the Academy of Marketing Science 44* (2016), 66–87. DOI: https://doi.org/10.1007/s11747-015-0441-x.

[22] Creswell, J. W. Chapter 18 - Mixed-Method Research: Introduction and Application. In *Handbook of Educational Policy*, G. J. Cizek, Ed., Educational Psychology. Academic Press, San Diego, Jan. 1999, pp. 455–472. https://www.sciencedirect.com/science/article/pii/B978012174698850045X [Accessed: 30.10.2024].

[23] Dabas, C., Kaur, P., Gulati, N., and Tilak, M. Analysis of Comments on Youtube Videos using Hadoop. In *2019 Fifth International Conference on Image Information Processing (ICIIP)* (Nov. 2019), pp. 353–358. ISSN: 2640-074X.

[24] Das, R., and Singh, T. D. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys 55*, 13s (2023), 1–38. DOI: https://doi.org/10.1145/3586075.

[25] DeepSeek-AI, Liu, A., Feng, B., Xue, B., et al. Deepseek-v3 technical report, 2024. https://arxiv.org/abs/2412.19437 [Accessed: 23.03.2025].

[26] Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR abs/1810.04805* (2018). http://arxiv.org/abs/1810.04805 [Accessed: 20.03.2025].

[27] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. https://arxiv.org/abs/1810.04805 [Accessed: 10.03.2025].

[28] Dudley, J. J., and Kristensson, P. O. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS) 8*, 2 (2018), 1–37. DOI: https://doi.org/10.1145/3185517.

[29] Durmus, E., Lovitt, L., Tamkin, A., Ritchie, S., Clark, J., and Ganguli, D. Measuring the persuasiveness of language models, 2024. https://www.anthropic.com/news/measuring-model-persuasiveness [Accessed: 03.03.2025].

[30] Eccles, D. W., and Arsal, G. The think aloud method: what is it and how do i use it? *Qualitative Research in Sport, Exercise and Health 9*, 4 (2017), 514–531.

[31] Eckifighter. Drachenlord-Beitrag | ZiB-Magazin(ORF), Aug. 2018. https://www.youtube.com/watch?v=3PbvS1x4gZM [Accessed: 29.10.2024].

[32] Epoch AI. data on machine learning hardware, 2024. https://epoch.ai/data/ai-benchmarking-dashboard [Accessed: 04.03.2025].

[33] Ericsson, K. A., and Simon, H. A. *Protocol Analysis: Verbal Reports As Data*, 2 ed. MIT Press, Cambridge, 1993.

[34] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. "i always assumed that i wasn't really that close to [her]" reasoning about invisible algorithms in news feeds. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (2015), pp. 153–162.

[35] Etikan, I. Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics 5* (2016), 1. DOI: https://doi.org/10.11648/j.ajtas.20160501.11.

[36] explainers-by-googlers. explainers-by-googlers/prompt-api, Oct. 2024. https://github.com/explainers-by-googlers/prompt-api [Accessed: 21.10.2024].

[37] Farias, D. H., and Rosso, P. Irony, sarcasm, and sentiment analysis. In *Sentiment analysis in social networks*. Elsevier, 2017, pp. 113–128. DOI: https://doi.org/10.1016/B978-0-12-804412-4.00007-3.

[38] Feuston, J. L., Taylor, A. S., and Piper, A. M. Conformity of eating disorders through content moderation. *Proceedings of the ACM on Human-Computer Interaction 4*, CSCW1 (2020), 1–28. DOI: https://doi.org/10.1145/3392845.

[39] Fitzgerald, J., and Sage, J. Shadowbans: Secret policies depriving sex workers of income and community. *Tits and Sass* (2019). https://titsandsass.com/shadowbans-secret-policies-depriving-sex-workers-of-income-and-community/ [Accessed: 23.03.2025].

[40] Franco, M., Gaggi, O., and Palazzi, C. E. Integrating content moderation systems with large language models. *ACM Transactions on the Web* (2024). DOI: https://doi.org/10.1145/3700789.

[41] Freitas dos Santos, T., Cranefield, S., Savarimuthu, B. T. R., Osman, N., and Schorlemmer, M. Cross-community adapter learning (cal) to understand the evolving meanings of norm violation. *International Joint Conferences on Artificial Intelligence* (2023). DOI: https://doi.org/10.24963/ijcai.2023/13.

[42] Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models, Sept. 2020. http://arxiv.org/abs/2009.11462 [Accessed: 31.10.2024].

[43] Gemma Team, Mesnard, T., Hardin, C., Dadashi, R., et al. Gemma, 2024. DOI: https://doi.org./10.34740/KAGGLE/M/3301.

[44] Gerrard, Y. Social media moderation: The best-kept secret in tech. In *The social media debate*. Routledge, 2022, pp. 77–95.

[45] GIFCT. Global internet forum to counter terrorism, 2019. https://perma.cc/44V5-554U [Accessed: 23.03.2025].

[46] Gillespie, T. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press, 2018.

[47] González-Ibánez, R., Muresan, S., and Wacholder, N. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (2011), pp. 581–586.

[48] Gorwa, R., Binns, R., and Katzenbach, C. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society 7*, 1 (Jan. 2020), 2053951719897945. DOI: https://doi.org/10.1177/2053951719897945.

[49] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models, 2024. https://arxiv.org/abs/2412.14093 [Accessed: 23.03.2025].

[50] Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., and Hu, H. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)* (2023), IEEE, pp. 1568–1573. DOI: https://doi.org/10.1109/ICMLA58977.2023.00237.

[51] Hamadaqa, M. H. M., Alnajjar, M., Ayyad, M. N., Al-Nakhal, M. A., Abunasser, B. S., and Abu-Naser, S. S. Leveraging artificial intelligence for strategic business decision-making: Opportunities and challenges. *International Journal of Academic Information Systems Research (IJAISR) 8*, 8 (2024), 16–23.

[52] Hinojosa Lee, M. C., Braet, J., and Springael, J. Performance metrics for multilabel emotion classification: Comparing micro, macro, and weighted f1-scores. *Applied Sciences 14*, 21 (2024). DOI: https://doi.org/10.3390/app14219863.

[53] Horta Ribeiro, M. *Content Moderation in Online Platforms.* PhD thesis, EPFL, 2024.

[54] Hugging Face. distilbert/distilbert-base-uncased-finetuned-sst-2-english ů Hugging Face, Jan. 2024. https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english [Accessed: 08.11.2024].

[55] Hugging Face. facebook/bart-large-mnli, Jan. 2024. https://huggingface.co/facebook/bart-large-mnli [Accessed: 23.03.2025].

[56] JAIS. Trusted Flagger bei YouTube So wirst du es, Apr. 2021. https://hateaid.org/trusted-flagger/ [Accessed: 21.10.2024].

[57] JHAVER, S., ZHANG, A. Q., CHEN, Q. Z., NATARAJAN, N., WANG, R., AND ZHANG, A. X. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction 7*, CSCW2 (2023), 1–33.

[58] JIANG, J. A., NIE, P., BRUBAKER, J. R., AND FIESLER, C. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction 30*, 1 (2023), 1–34. DOI: https://doi.org/10.1145/3534929.

[59] JIGSAW. Unintended Bias and Identity Terms, Oct. 2021. https://medium.com/jigsaw/unintended-bias-and-names-of-frequently-targeted-groups-8e0b81f80a23 [Accessed: 30.10.2024].

[60] JIGSAW. Scaling machine learning fairness with societal context, Dec. 2022. https://medium.com/jigsaw/scaling-machine-learning-fairness-with-societal-context-be73d4ad38e2 [Accessed: 30.10.2024].

[61] JIGSAW. Announcing Experimental Bridging Attributes in Perspective API, Apr. 2024. https://medium.com/jigsaw/announcing-experimental-bridging-attributes-in-perspective-api-578a9d59ac37 [Accessed: 30.10.2024].

[62] JOSHI, A., KALE, S., CHANDEL, S., AND PAL, D. K. Likert scale: Explored and explained. *British journal of applied science & technology 7*, 4 (2015), 396. DOI: https://doi.org/10.9734/BJAST/2015/14975.

[63] JOSHUA WEISSMAN. Can You Trust A Skinny Chef (How I Lost 60 Lbs), Sept. 2024. https://www.youtube.com/watch?v=4-XPa09H1Xs [Accessed: 23.03.2025].

[64] KAGGWA, S., ELEOGU, T. F., OKONKWO, F., FARAYOLA, O. A., UWAOMA, P. U., AND AKINOSO, A. Ai in decision making: transforming business strategies. *International Journal of Research and Scientific Innovation 10*, 12 (2024), 423–444. DOI: https://doi.org/10.51244/IJRSI.2023.1012032.

[65] KAUR, R., AND KAUTISH, S. Multimodal sentiment analysis: A survey and comparison. *Research anthology on implementing sentiment analysis across multiple disciplines* (2022), 1846–1870. DOI: https://doi.org/10.4018/978-1-6684-6303-1.ch098.

[66] KAYYALI, D. Meta's content moderation changes are going to have a real world impact. it's not going to be good. *Tech Policy Press* (2025). https://www.techpolicy.press/metas-content-moderation-changes-are-going-to-have-a-real-world-impact-its-not-going-to-be-good/ [Accessed: 13.03.2025].

[67] KLONICK, K. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev. 131* (2017), 1598.

[68] Kumar, A., Singhal, S., and Sharma, A. Analyzing sentiments of youtube comments using textblob and machine learning algorithms. In *International Conference on Information Technology* (2024), Springer, pp. 99–110.

[69] Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y., and Tan, C. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022), pp. 1–18. DOI: https://doi.org/10.1145/3491102.3501999.

[70] Lin, Z. J., Jung, J., Goel, S., and Skeem, J. The limits of human predictions of recidivism. *Science advances 6*, 7 (2020), eaaz0652. DOI: https://doi.org/10.1126/sciadv.aaz0652.

[71] Liu, C., Fang, F., Lin, X., Cai, T., Tan, X., Liu, J., and Lu, X. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience 2*, 4 (2021), 246–252. DOI: https://doi.org/10.1016/j.jnlssr.2021.10.003.

[72] Ma, R., and Kou, Y. "I'm not sure what difference is between their content and mine, other than the person itself" A Study of Fairness Perception of Content Moderation on YouTube. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW2 (2022), 1–28. DOI: https://doi.org/10.1145/3555150.

[73] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Portland, Oregon, USA, June 2011), Association for Computational Linguistics, pp. 142–150. http://www.aclweb.org/anthology/P11-1015 [Accessed: 23.03.2025].

[74] Machine Learning Compilation. WebLLM. https://webllm.mlc.ai/ [Accessed: 21.10.2024].

[75] Marks, S., Treutlein, J., Bricken, T., Lindsey, J., Marcus, J., Mishra-Sharma, S., Ziegler, D., Ameisen, E., Batson, J., Belonax, T., et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965* (2025).

[76] Matsakis, L., and Martineau, P. Coronavirus disrupts social media's first line of defense. *WIRED* (2020). https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/ [Accessed: 23.03.2025].

[77] Meghana, K. Artificial Intelligence and Sentiment Analysis in YouTube Comments: A Comprehensive Overview. In *2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)* (Jan. 2024), pp. 1565–1572. DOI: https://doi.org/10.1109/IDCIoT59759.2024.10467782.

[78] Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., and Wihbey, J. P. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology 73*, 10 (2022), 1365–1386. DOI: https://doi.org/10.1002/asi.24637.

[79] Mukherjee, S., Coulter, M., Dang, S., and Mukherjee, S. Twitter blue bird has flown as Musk says X logo is here. *Reuters* (July 2023). https://www.reuters.com/technology/bird-has-flown-musk-twitter-ceo-yaccarino-say-x-logo-is-here-2023-07-24/ [Accessed; 04.03.2025].

[80] N, P., G, N., U, B. K., Vardhan B, H., and L, T. K. DeTox: A WebApp for Toxic Comment Detection and Moderation. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies* (Mar. 2024), pp. 1–5. DOI: https://doi.org/10.1109/TQCEBT59414.2024.10545229.

[81] Nagulendra, S., and Vassileva, J. Providing awareness, explanation and control of personalized filtering in a social networking site. *Information Systems Frontiers 18* (2016), 145–158. DOI: https://doi.org/10.1007/s10796-015-9577-y.

[82] Nurik, C. Men are scum: Self-regulation, hate speech, and gender-based censorship on Facebook. *International Journal of Communication 13* (2019), 21.

[83] Ollama. Get up and running with large language models, 2025. https://ollama.com/ [Accessed: 28.02.2025].

[84] OpenAI. Chatgpt, 2022. https://openai.com/index/chatgpt/ [Accessed: 09.10.2024].

[85] OpenAI. openai/whisper, Feb. 2025. https://github.com/openai/whisper [Accessed: 14.02.2025].

[86] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., et al. Training language models to follow instructions with human feedback, Mar. 2022. http://arxiv.org/abs/2203.02155 [Accessed: 30.10.2024].

[87] Padma, T., Visweshvar, R., Tamilarasan, K., and Bhadrinath, C. J. Dynamic YouTube Comment Sentiment Analysis with Supervised Fine-Tuned BERT. In *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)* (Apr. 2024), pp. 663–669. DOI: htpps://doi.org/10.1109/ICC-ROBINS60238.2024.10533926.

[88] Pan, C. A., Yakhmi, S., Iyer, T. P., Strasnick, E., Zhang, A. X., and Bernstein, M. S. Comparing the perceived legitimacy of content moderation processes: Contractors, algorithms, expert panels, and digital juries. *Proceedings of the ACM on Human-Computer Interaction 6*, CSCW1 (2022), 1–31. DOI: https://doi.org/10.1145/3512929.

[89] PERSPECTIVE. About the API - Attributes and Languages. https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US [Accessed: 22.10.2024].

[90] PERSPECTIVE. Perspective API. https://perspectiveapi.com/ [Accessed: 21.10.2024].

[91] RAUFEISEN, X., WULF, L., KÖCHER, S., FAUPEL, U., AND HOLZMÜLLER, H. H. Spillover effects in marketing: integrating core research domains. *AMS Review 9* (2019), 249–267.

[92] RENDALKAR, S., AND CHANDANKHEDE, C. Sarcasm detection of online comments using emotion detection. In *2018 International conference on inventive research in computing applications (ICIRCA)* (2018), IEEE, pp. 1244–1249. DOI: https://doi.org/10.1109/ICIRCA.2018.8597368.

[93] REYES, A., ROSSO, P., AND BUSCALDI, D. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering 74* (2012), 1–12. DOI: https://doi.org/10.1016/j.datak.2012.02.005.

[94] RIORDAN, M. A. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology 36*, 5 (2017), 549–567. DOI: https://doi.org/10.1177/0261927X17704238.

[95] ROBERTS, S. Commercial content moderation: Digital laborers' dirty work. *Media Studies Publications, 12* (2016).

[96] SALTZ, E., JALAN, Z., AND ACOSTA, T. Re-Ranking News Comments by Constructiveness and Curiosity Significantly Increases Perceived Respect, Trustworthiness, and Interest, Apr. 2024. http://arxiv.org/abs/2404.05429 [Accessed: 30.10.2024].

[97] SANH, V., DEBUT, L., CHAUMOND, J., AND WOLF, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108.

[98] SHARMA, M., TONG, M., MU, J., ET AL. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025. https://arxiv.org/abs/2501.18837 [Accessed: 23.03.2025].

[99] SHETH, B., AND MAES, P. Evolving agents for personalized information filtering. In *Proceedings of 9th IEEE conference on artificial intelligence for applications* (1993), IEEE, pp. 345–352. DOI: https://doi.org/10.1109/CAIA.1993.366590.

[100] SHIHA, M., AND AYVAZ, S. The effects of emoji in sentiment analysis. *Int. J. Comput. Electr. Eng.(IJCEE.) 9*, 1 (2017), 360–369.

[101] Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies 146* (2021), 102551.

[102] Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 2003, pp. 364–371.

[103] Shulner-Tal, A., Kuflik, T., and Kliger, D. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users perceptions of fairness toward an algorithmic system. *Ethics and Information Technology 24*, 1 (Jan. 2022), 2.

[104] similarweb. Top Websites Ranking - Most Visited Websites in September 2024 worldwide. https://www.similarweb.com/top-websites/ [Accessed: 21.10.2024].

[105] sinders, c. Toxicity and Tone Are Not The Same Thing: analyzing the new Google API on toxicity, PerspectiveAPI., Feb. 2017. https://medium.com/@carolinesinders/toxicity-and-tone-are-not-the-same-thing-analyzing-the-new-google-api-on-toxicity-perspectiveapi-14abe4e728b3 [Accessed: 29.10.2024].

[106] Smith, C. E., Yu, B., Srivastava, A., Halfaker, A., Terveen, L., and Zhu, H. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–14. DOI: https://doi.org/10.1145/3313831.3376783.

[107] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington, USA, Oct. 2013), Association for Computational Linguistics, pp. 1631–1642. https://www.aclweb.org/anthology/D13-1170 [Accessed: 20.03.2025].

[108] Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. A survey of multimodal sentiment analysis. *Image and Vision Computing 65* (2017), 3–14. DOI: https://doi.org/10.1016/j.imavis.2017.08.003.

[109] Statcounter. Browser Market Share Worldwide (feb 2024 - feb 2025). https://gs.statcounter.com/browser-market-share [Accessed: 03.03.2025].

[110] stern TV. Der Drachenlord und seine Hater: Chronologie einer Eskalation, Nov. 2021. https://www.youtube.com/watch?v=3rw0NmKw7as [Accessed: 23.03.2025].

[111] Struthoff, N. Drachenlord muss nicht ins Gefängnis – Staatsanwaltschaft lenkt ein, June 2022. https://www.ingame.de/news/streaming/drachenlord-urteil-bewaehrung-gerichtsverfahren-gefaengnis-knast-hater-entscheidung-revision-staatsanwaltschaft-nuernberg-91584713.html [Accessed: 01.11.2024].

[112] Subramanian, J., Sridharan, V., Shu, K., and Liu, H. Exploiting emojis for sarcasm detection. In *Social, Cultural, and Behavioral Modeling: 12th International Conference, SBP-BRiMS 2019, Washington, DC, USA, July 9–12, 2019, Proceedings 12* (2019), Springer, pp. 70–80. DOI: https://doi.org/10.1007/978-3-030-21741-9_8.

[113] Templeton, A., Conerly, T., Marcus, J., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread* (2024). https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html [Accessed: 23.03.2025].

[114] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., et al. LLaMA: Open and Efficient Foundation Language Models, Feb. 2023. http://arxiv.org/abs/2302.13971 [Accessed: 30.10.2024].

[115] Vaccaro, K., Xiao, Z., Hamilton, K., and Karahalios, K. Contestability for content moderation. *Proceedings of the ACM on human-computer interaction 5*, CSCW2 (2021), 1–28. DOI: https://doi.org/10.1145/3476059.

[116] Van Teijlingen, E. R., Rennie, A.-M., Hundley, V., and Graham, W. The importance of conducting and reporting pilot studies: the example of the scottish births survey. *Journal of advanced nursing 34*, 3 (2001), 289–295.

[117] White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., et al. Livebench: A challenging, contamination-free LLM benchmark. In *The Thirteenth International Conference on Learning Representations* (2025).

[118] Wihbey, J., Kopec, M., and Sandler, R. Informational quality labeling on social media: in defense of a social epistemology strategy. *Special Issue:'Social Media Governance 23* (2021).

[119] Wilson, R. A., and Land, M. K. Hate speech on social media: Content moderation in context. *Connecticut Law Review 52* (2020), 1029.

[120] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L.-P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems 28*, 3 (2013), 46–53. DOI: https://doi.org/10.1109/MIS.2013.34.

[121] Xu, T., and Baghaei, S. Reshaping the future of sports with artificial intelligence: Challenges and opportunities in performance enhancement, fan engagement, and strategic decision-making. *Engineering Applications of Artificial Intelligence 142* (2025), 109912. DOI: https://doi.org/10.1016/j.engappai.2024.109912.

[122] Y-Kollektiv. Drachenlord und seine Hater - Hass ist ihr Hobby - RABIAT!, June 2018. https://www.youtube.com/watch?v=zu9KtSvFGMI [Accessed: 29.10.2024].

[123] Yang, R., Singh, S., Cao, P., Chi, E., and Fu, B. Video Watch Time and Comment Sentiment: Experiences from YouTube. In *2016 Fourth IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)* (Oct. 2016), pp. 26–28. Doi: https://doi.org/10.1109/HotWeb.2016.13.

[124] YouTube Hilfe. Das YouTube Priority Flagger Program - YouTube-Hilfe. https://support.google.com/youtube/answer/7554338 [Accessed: 21.10.2024].

[125] Zhan, X., Goyal, A., Chen, Y., Chandrasekharan, E., and Saha, K. SLM-Mod: Small Language Models Surpass LLMs at Content Moderation, Feb. 2025. http://arxiv.org/abs/2410.13155 [Accessed: 28.02.2025].

[126] Zheng, A. H. ASCII Smuggling and Hidden Prompt Instructions Attack, Aug. 2024. https://securaize.substack.com/p/ascii-smuggling-and-hidden-prompt [Accessed: 22.03.2025].

[127] Zhu, H., Yu, B., Halfaker, A., and Terveen, L. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on human-computer interaction 2*, CSCW (2018), 1–23. DOI: https://doi.org/10.1145/3274463.