CommentSense* Meetup Edition

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

Marc Kroll

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

Available Models[1]

April, 2024
Llama-3-8B & 70B
Phi-3-Mini

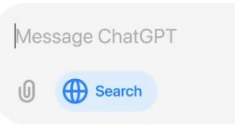
May, 2024
Phi-3 small/medium

June, 2024
Gemma 2

December, 2024
Phi-4

September, 2024
o1-preview & o1-mini [2]

October, 2024
ChatGPT Search [2]



[1] https://github.com/eugeneyan/open-llms
[2] https://openai.com/o1/, https://openai.com/index/introducing-chatgpt-search/

“deepseek moment”

Available Models

January, 2025
mistral-small-3[1]

March, 2025
Gemma 3[2]
mistral-small-3.1[1]

\+ DeepSeek V3 & R1
Qwen 2.5 & Qwen 3

Alignment

Reasoning models don't always say what they think

3 Apr 2025

Read the paper

Since late last year, "reasoning models" have been everywhere. These are AI models—such as Claude 3.7 Sonnet—that *show their working*: as well as their eventual answer, you can read the (often fascinating and convoluted) way that they got there, in what's called their "Chain-of-Thought".

As well as helping reasoning models work their way through more difficult problems, the Chain-of-Thought has been a boon for AI safety researchers. That's because we can (among other things) check for things the model says in its Chain-of-Thought that go unsaid in its output, which can help us spot undesirable behaviours like deception.

But if we want to use the Chain-of-Thought for alignment purposes, there's a crucial question: can we actually *trust* what models say in their Chain-of-Thought?

In a perfect world, everything in the Chain-of-Thought would be both understandable to the reader, and it would be *faithful*—it would be a true description of exactly what the model was thinking as it reached its answer.

But we're not in a perfect world. We can't be certain of either the

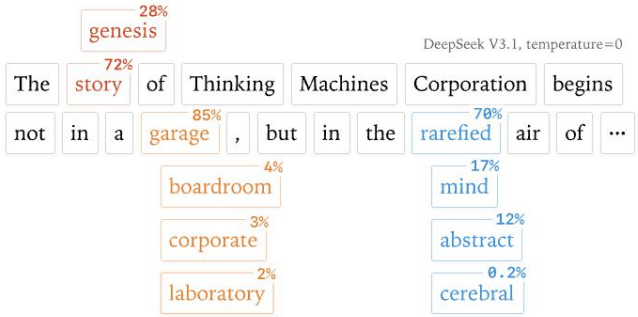
[1] https://mistral.ai/news/mistral-small-3, https://mistral.ai/news/mistral-small-3-1
[2] https://blog.google/technology/developers/gemma-3/
[3] https://arxiv.org/pdf/2505.05410

September 2025

Defeating Nondeterminism in LLM Inference

Horace He in collaboration with others at Thinking Machines

Sep 10, 2025



Reproducibility is a bedrock of scientific progress. However, it's remarkably difficult to get reproducible results out of large language models.

Small Language Models are the Future of Agentic AI

Peter Belcak[1] Greg Heinrich[1] Shizhe Diao[1] Yonggan Fu[1] Xin Dong[1]
Saurav Muralidharan[1] Yingyan Celine Lin[1,2] Pavlo Molchanov[1]
[1]NVIDIA Research [2]Georgia Institute of Technology
agents-research@nvidia.com

Abstract

Large language models (LLMs) are often praised for exhibiting near-human performance on a wide range of tasks and valued for their ability to hold a general conversation. The rise of agentic AI systems is, however, ushering in a mass of applications in which language models perform a small number of specialized tasks repetitively and with little variation.

Here we lay out the position that small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI*. Our argumentation is grounded in the current level of capabilities exhibited by SLMs, the common architectures of agentic systems, and the economy of LM deployment. We further argue that in situations where general-purpose conversational abilities are essential, heterogeneous agentic systems (i.e., agents invoking multiple different models) are the natural choice. We discuss the potential barriers for the adoption of SLMs in agentic systems and outline a general LLM-to-SLM agent conversion algorithm.

Our position[1], formulated as a value statement, highlights the significance of the operational and economic impact even a partial shift from LLMs to SLMs is to have on the AI agent industry. We aim to stimulate the discussion on the effective use of AI resources and hope to advance the efforts to lower the costs of AI of the present day. Calling for both contributions to and critique of our position, we commit to publishing all such correspondence at `research.nvidia.com/labs/lpr/slm-agents`.

1 Introduction

The deployment of agentic artificial intelligence is on a meteoric rise. Recent surveys show that more than a half of large IT enterprises are actively using AI agents, with 21% having adopted just within the last year [14]. Aside from the users, markets also see substantial economic value in AI agents: As of late 2024, the agentic AI sector had seen more than USD 2bn in startup funding, was valued at USD 5.2bn, and was expected to grow to nearly USD 200bn by 2034 [46, 51]. Put plainly, there is a growing expectation that AI agents will play a substantial role in the modern economy.

The core components powering most modern AI agents are (very) large language models [52, 48]. It is the LLMs that provide the foundational intelligence that enables agents to make strategic decisions about when and how to use available tools, control the flow of operations needed to complete tasks, and, if necessary, to break down complex tasks into manageable subtasks and to perform reasoning for action planning and problem-solving [52, 17]. A typical AI agent then simply communicates with a chosen LLM API endpoint by making requests to centralized cloud infrastructure that hosts these models [52].

[1]The views and positions expressed in this paper are those of the authors and do not necessarily reflect the views or positions of any entities they represent.

Preprint. Under review.

[1] https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/
[2] https://arxiv.org/pdf/2506.02153

CommentSense* Meetup Edition

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

Marc Kroll

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

[1] Joshua Weissman. Can You Trust A Skinny Chef (How I Lost 60 Lbs), Sept. 2024. https://www.youtube.com/watch?v=4-XPaO9H1Xs

@russellyork5853
Why do u look so fat????

@Tai-oLo
you so fat

@anthonysmith1591
Grossly obese person telling you what food they don't like is hilarious.

@dannyespinoza3836
Chill fatty

@lnsayenign9422
Bro you're getting fat!

@nestorrfortuna1
Getting fat

@TheDancerPL
bro u got fat:(

@greeneffectltd
Joshua Fatman is your new name fatso!

@coltonharris1526
Josh is getting fat

@Neobanned
Josh, hows the weight loss going you phat

@trickyd499 1 year ago
fatty fatty bum bum

@jernigan007
you got fat

@beaualoevv
notice how he gained weight lol. he should workout

@MrKojotie
Aloha, now you're fat :P

@Mkantae
blud is fat now.

@kucci1769 1 year ago
josh ur getting fat.

@roadrunner1095
With every video you are getting fatter... Do some exercise man.

[1] Joshua Weissman. Can You Trust A Skinny Chef (How I Lost 60 Lbs), Sept. 2024. https://www.youtube.com/watch?v=4-XPaO9H1Xs

[1] Joshua Weissman. Can You Trust A Skinny Chef (How I Lost 60 Lbs), Sept. 2024. https://www.youtube.com/watch?v=4-XPaO9H1Xs

Model comparison for the term: “Ozempic face”

- Severe Toxicity
- Identity attack
- Insult
- Threat
- Profanity
- Sexually explicit

Llama 3.2 3b [4]

SENTIMENT:
NEGATIVE

REASONING:
Phrases like Ozempic face are often used to criticize the appearance resulting from weight loss medications, carrying a negative connotation.

Perspective API [2]

Toxicity: 16,76%
Insult: 8.36%

DistilBERT-sst2 [3]

POSITIVE: 57,89%

BERT-base model fine-tuned for doing binary (positive / negative) sentiment analysis

[2] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages
[3] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english
[4] https://huggingface.co/meta-llama/Llama-3.2-3B

Are Small Language Models better in sentiment analysis than BERT-based models?

Hypothesis 1: SLMs outperform BERT-based models due to their contextual knowledge.

Hypothesis 2: Users prefer SLMs due to their reasoning capabilities.

High Level	Specific	Trends
Content Moderation	Hate Speech Detection	Human-in-the-Loop Systems
Fairness in Content Classification	Irony, Sarcasm, Emoji Detection	Text Classification with Language Models
Value-Sensitive Algorithms	Personalized Filtering	Multimodal Sentiment Analysis

Are Small Language Models better in sentiment analysis than BERT-based models?

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 language model in doing sentiment analysis of YouTube comments?

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?

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?

real model evaluation

Step 1

User Preference

6 Participants
1h Remote-Session
Interactive Prototype
Semi-structured interview
Note taking

Step 2

Model Evaluation

19 Small Language Models*
1 fine-tuned BERT Model
Human ground truth
YouTube dataset
Binary classification
Three-way classification
Performance metrics
Macro F1-Score

Step 3

User Evaluation

7 Participants
1 ½ h in-person sessions
Surveys
User tasks + think aloud
Observation
Note taking
Semi-structured interview
Thematic analysis
Human ground truth
YouTube dataset

Mixed-Method approach [5]

Quantitative methods:
Performance metrics, Macro F1-Score, Likert-scales in surveys

Qualitative methods:
Semi-structured interviews, think aloud, observation, note taking, thematic analysis, open questions in surveys

* Models had to be below 22b parameters, multilingual, public and for free without any constraints (sign-up, gated, experimental, etc)

[5] Creswell, J. W. Chapter 18 - Mixed-Method Research: Introduction and Application.

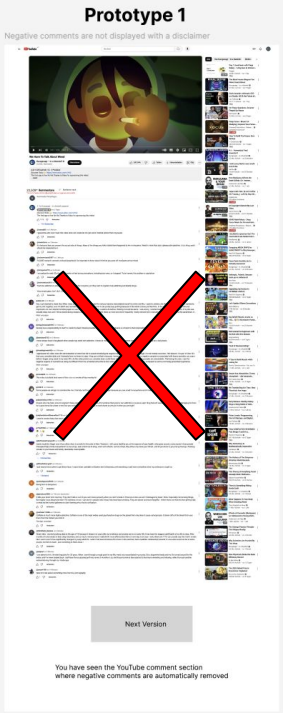
Feedback

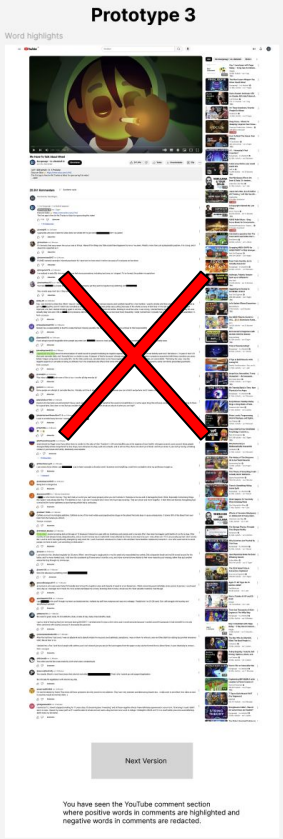
Interactive prototype

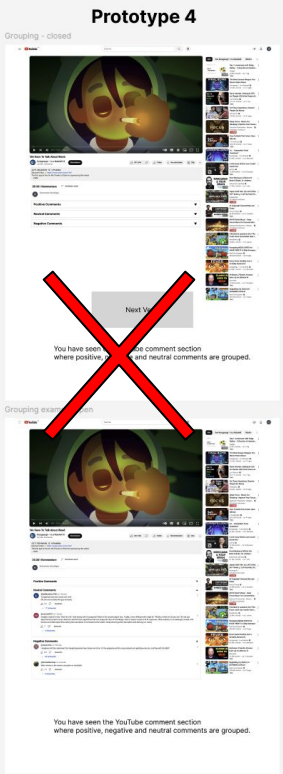
Don’t group!
Don’t hide anything!
Don’t change the appearance!

Ich möchte eher die negativen auch.

Kann ich bestimmen was raus soll?







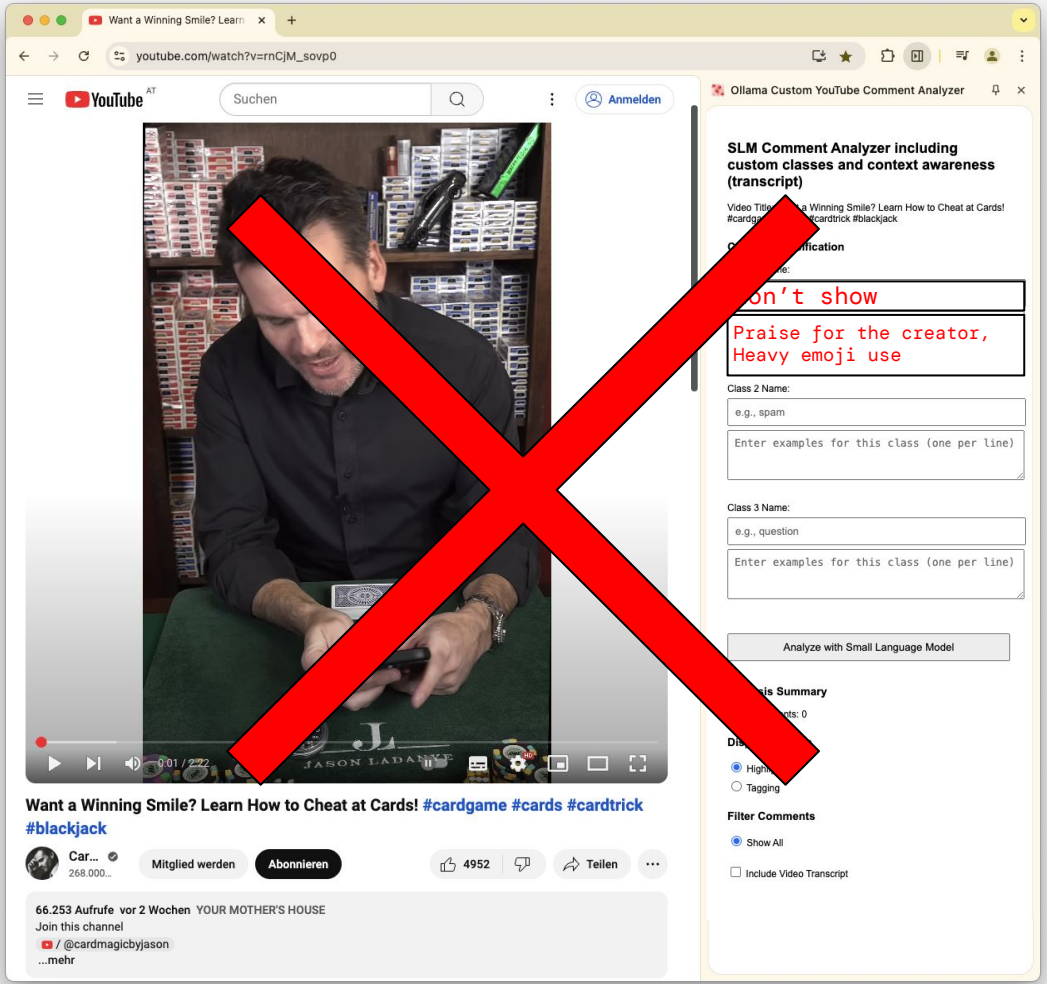
Pilot 1

Custom Classification

Participants struggled defining classes and examples for classification.

was soll ich da jetzt reinschreiben?

Einfach irgendwas?



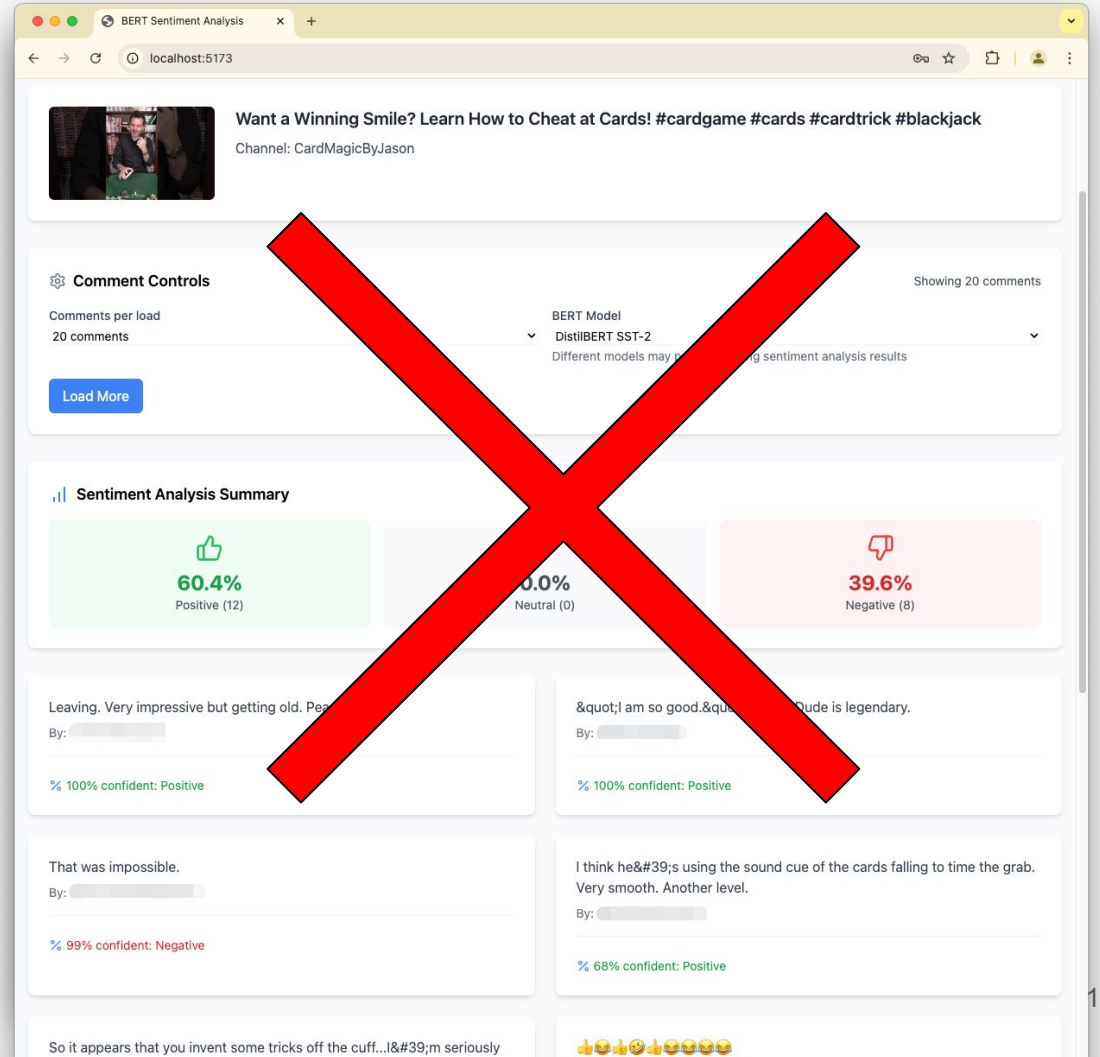
Pilot 2

Extension vs Website

Too many options.
Task sheet to complex.
Cognitive load was to high.
Pilot had to be stopped early.

Muss ich wirklich alle Optionen testen?

Simplify training, user tasks and extension.



Are Small Language Models better in sentiment analysis than BERT-based models?

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 language model in doing sentiment analysis of YouTube comments?

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?

Simplicity and maximum transparency. No reordering, grouping or highlighting. Negative content is preferred to be displayed instead of hidden.

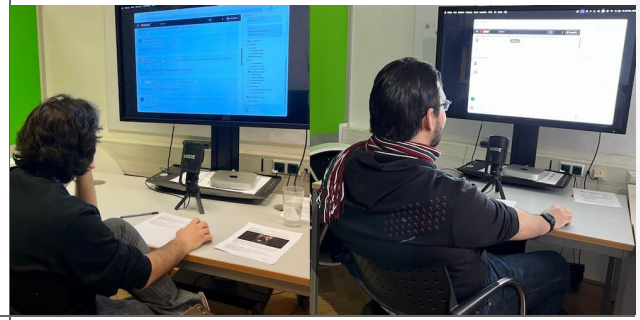
RQ3: (Bridging technical and human-centered)
How does the use of small language models compare to a state-of-the-art BERT-based model and influence users' perceived usefulness of sentiment analysis systems for YouTube comments?

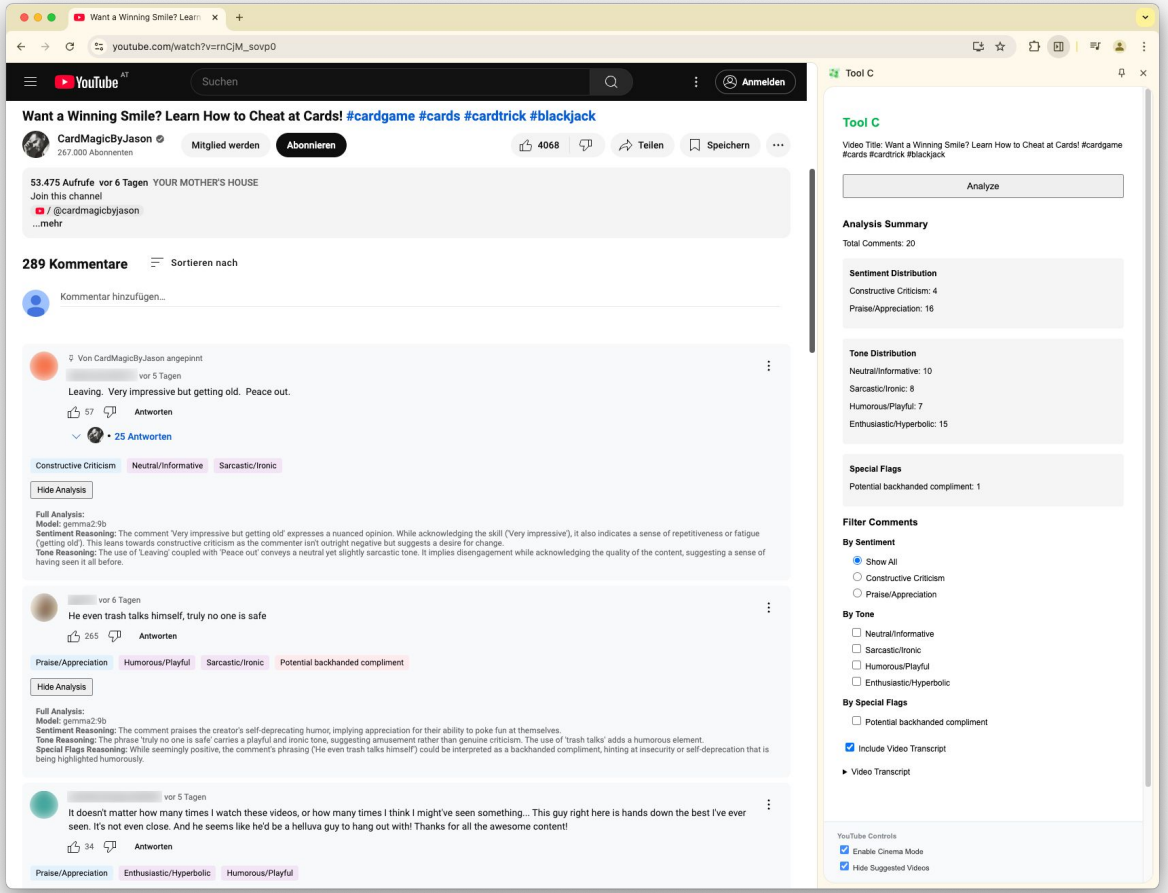
User Evaluation (n=7)

3 Tools

Study setting

Training video
Pre-task survey
User tasks + think aloud
Post-task survey
Interview





Note: All data was processed locally.

Final Extension:

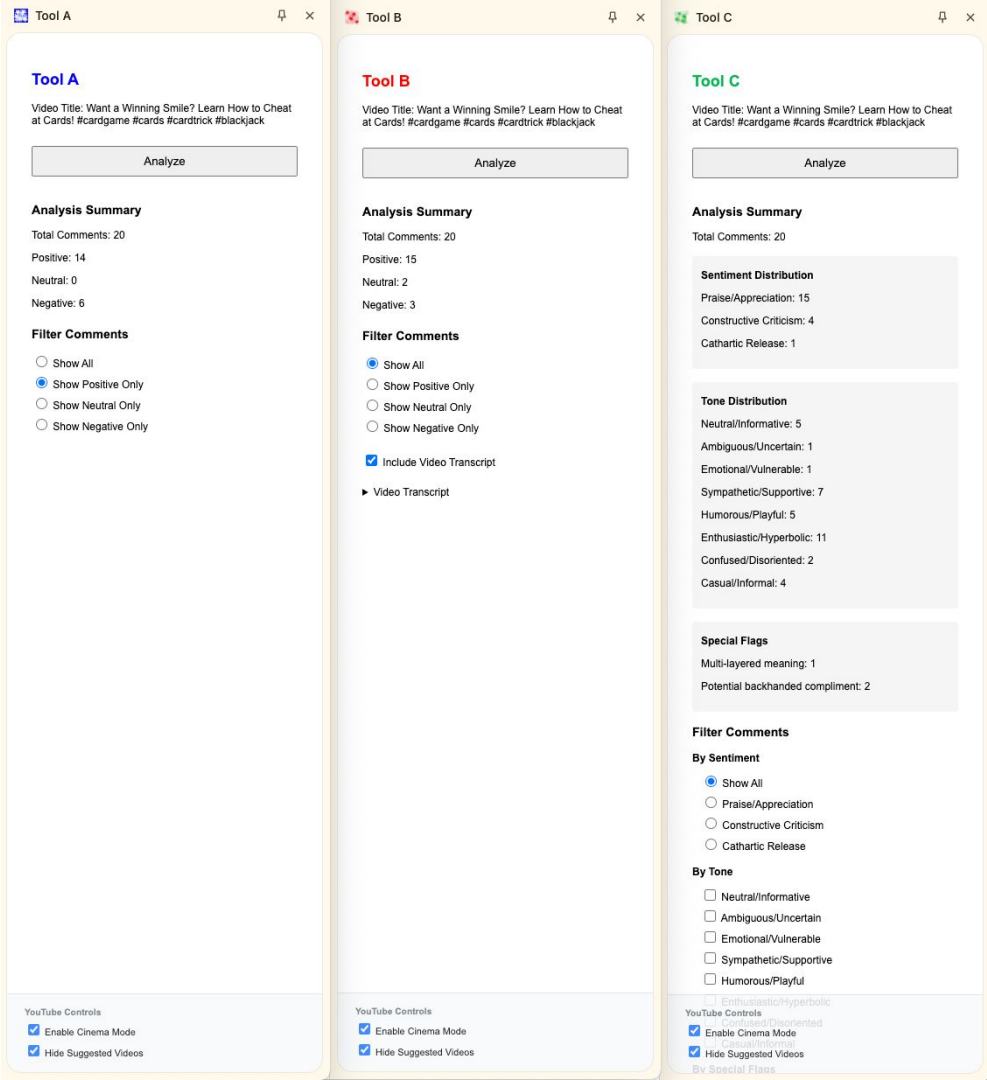
Side panels

Tool A:
- Distilbert-sst2

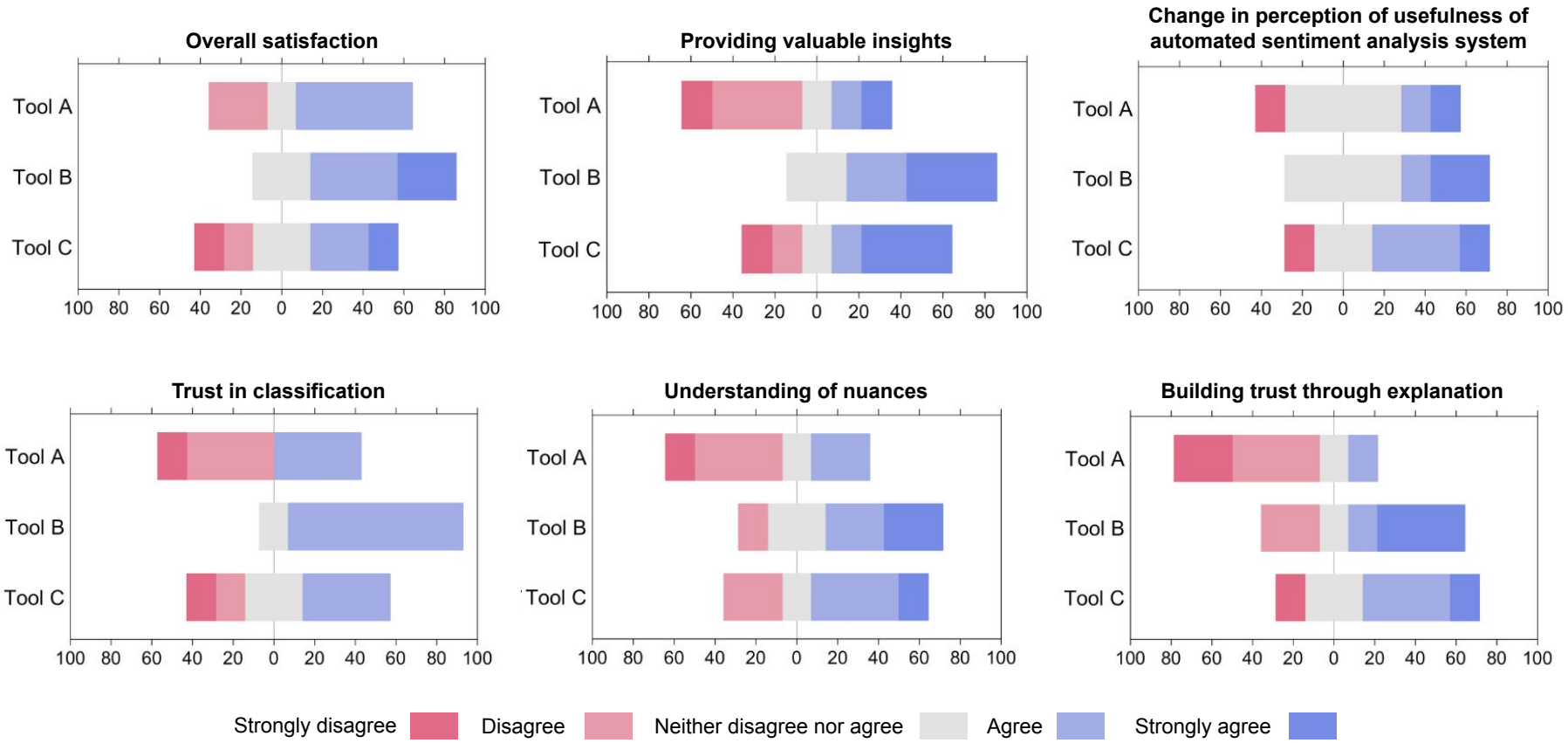
Tool B & Tool C:
- Mistral-small 3

Summary and Filter Options.

Context of the Video (transcript) can be included.

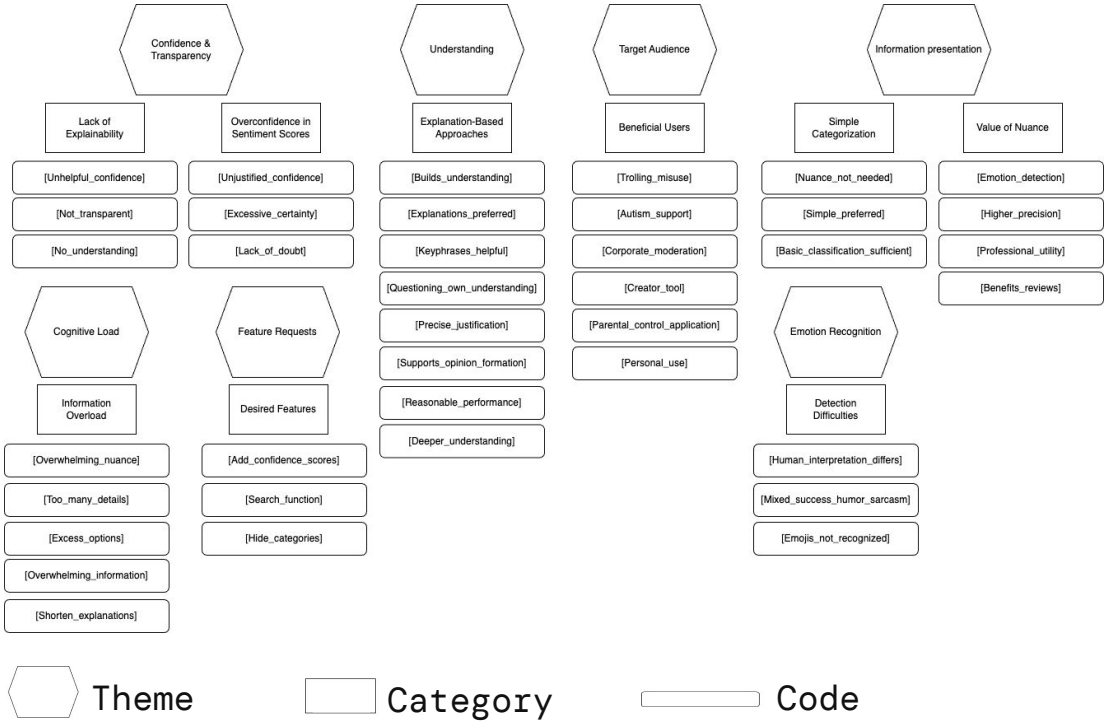


User Evaluation: post-task survey



Note: all X-axis in percent (%). Label was removed to enhance readability in the presentation.
Tool A: BERT-based — Tool B & C: SLM-based.

Thematic analysis [6]



Finding 1

Nuanced and detailed classifications were not received well by most participants due to high cognitive load. Simplicity is preferred.

Finding 2

Explanations for classifications were preferred from all participants compared to confidence scores.

Finding 3

The extensions are suggested to be useful for parental control, YouTube creators or to support people with autism to better understand emotional cues in language.

[6] Braun, V., and Clarke, V. Conceptual and design thinking for thematic analysis. Qualitative psychology 9, 1 (2022), 3

Are Small Language Models better in sentiment analysis than BERT-based models?

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 language model in doing sentiment analysis of YouTube comments?

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?

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?

Simplicity and maximum transparency. No reordering, grouping or highlighting. Negative content is preferred to be displayed instead of hidden.

The SLM based approach positively changed the perceived usefulness of a sentiment analysis system due to explanations. The BERT-based approach was perceived neutral.

Macro F1-Score

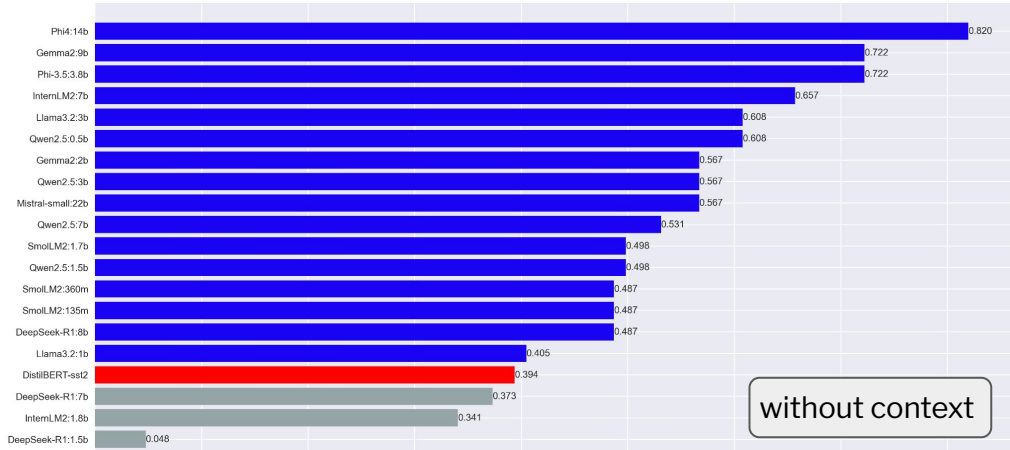
Binary Classification

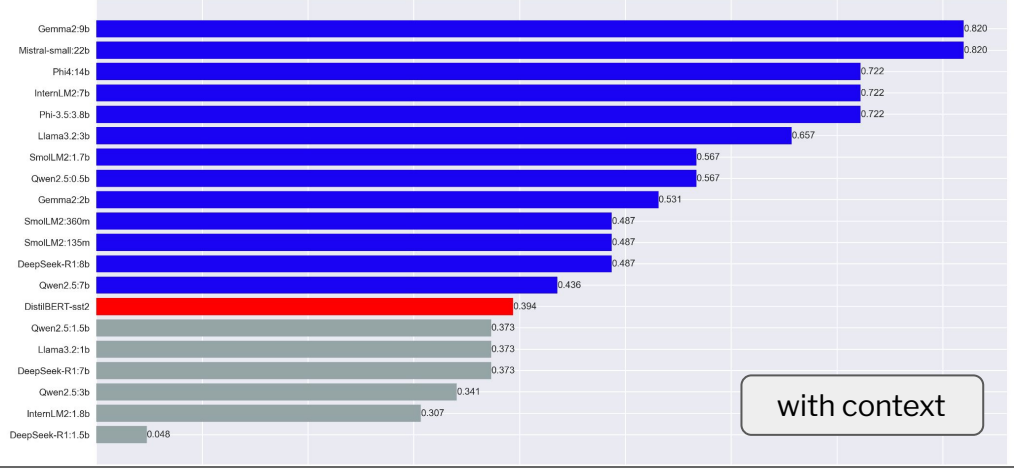
Measuring correct classification of positive, negative against human ground truth.

Best:
`phi4:14b, gemma2:9b, mistral-small:22b`

Worst:
`internlm2:1.8b, deepseek-r1:1.5b`

Macro F1-Score





Notes: Model evaluation was performed with fixed seed and temperature.
Human annotation was adjusted for binary classification by the researcher.
`DistilBERT-sst2` was not capable of taking the evaluated context into account (token limit)

Macro F1-Score

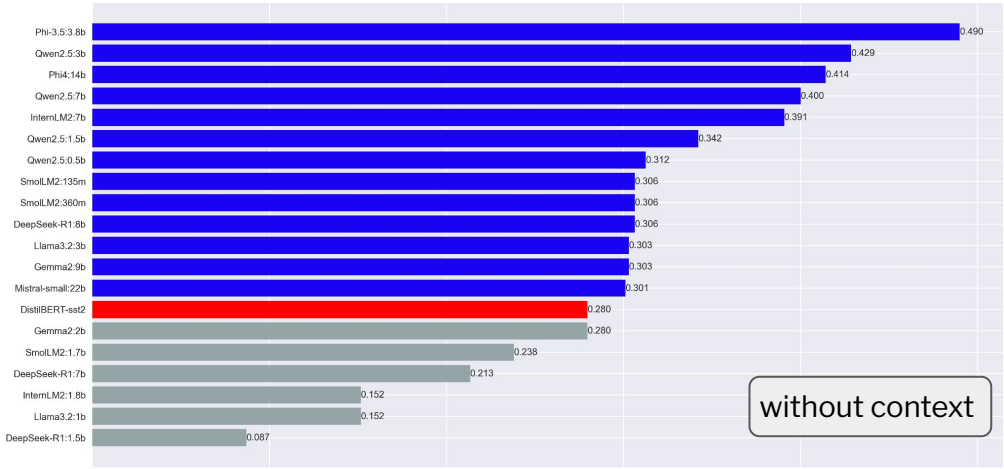
Three-way Classification*

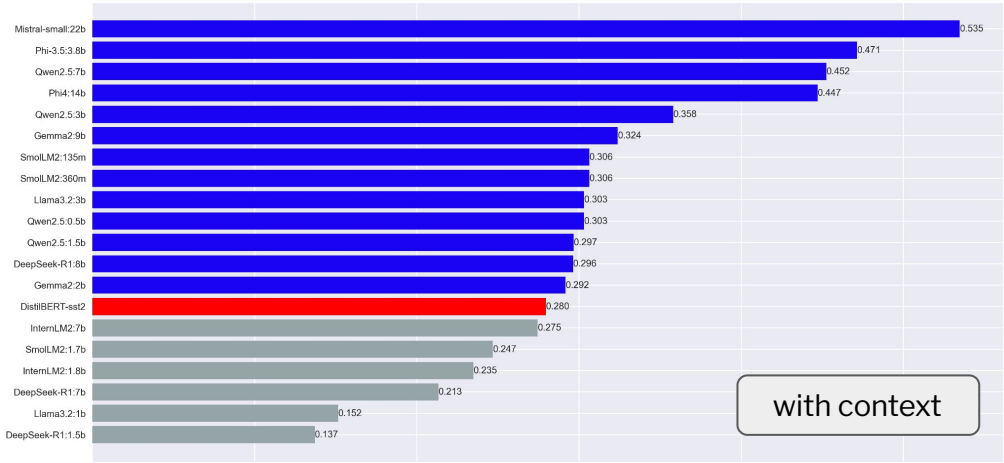
Measuring correct classification of positive, negative, neutral against human ground truth.

Best:
`phi3.5:3.8b, mistral-small:22b`

Worst:
`llama3.2:1b, deepseek-r1:1.5b`

Macro F1-Score





Note: `DistilBERT-sst2` is only able to do binary classification
*`Users stated: positive/negative is not enough. Therefore “neutral” was introduced as class.`

Are Small Language Models better in sentiment analysis than BERT-based models?

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 language model in doing sentiment analysis of YouTube comments?

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?

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?

Over 50% of all evaluated SLMs outperformed DistilBERT-sst2 based on Macro F1-Scores against the human ground truth for a dedicated dataset.

Simplicity and maximum transparency. No reordering, grouping or highlighting. Negative content is preferred to be displayed instead of hidden.

The SLM based approach positively changed the perceived usefulness of a sentiment analysis system due to explanations. The BERT-based approach was perceived neutral.

For a sentiment analysis tool ..

.. to be perceived as trustworthy and useful, transparency in its classifications is essential, along with a simple and unobtrusive display method.

Concerns included over-filtering, bias, free speech, and privacy, with a focus on transparency. The tool must balance accuracy, control, and transparency for acceptance.

Extensions using SLMs ranked highest for transparency, usefulness, and explanation quality, enhancing user understanding of YouTube comment sentiment and driving acceptance.

Explanations for classification results were preferred by every participant instead of confidence scores.

Nuanced and detailed classifications were not received well by most participants.

Traditional sentiment analysis approaches should be reconsidered.

Best performing SLMs* sorted by parameter size: `phi3.5:3.8b` - `gemma2:9b` - `phi4:14b` - `mistral-small:22b`
SLMs from the phi, gemma and mistral series outperformed DistilBERT-sst2 in all tests.

* General purpose SLMs are evaluated without any dedicated fine-tuning for sentiment analysis .
Note: always the latest release of available SLMs should be evaluated.

Sources

[1] Joshua Weissman. Can You Trust A Skinny Chef (How I Lost 60 Lbs), Sept. 2024. https://www.youtube.com/watch?v=4-XPaO9H1Xs

[2] https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

[3] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

[4] https://huggingface.co/meta-llama/Llama-3.2-3B

[5] 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.

[6] Braun, V., and Clarke, V. Conceptual and design thinking for thematic analysis. Qualitative psychology 9, 1 (2022), 3

A custom Chrome extension enabling users to perform sentiment analysis while staying on YouTube, supporting the field of user-centered content moderation while comparing a BERT-based model with SLMs with a focus on the user as the central recipient of the results.

(Technical)
An open-source 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.

Thanks.

26.03.2025

'V 'Perspective Team' via pe... Montag
An: undisclosed-recip ... & 1 weitere >
Antwort an: Perspective Team >

[Perspective API] Customizable attributes beta in Perspective API

Dear Perspective users and friends,

We recently announced customizable attributes in Perspective API. Users can provide their community's rules in plain English, as well as example comments and decisions, and receive a score that reflects whether a comment complies with their rules. This capability, powered by the latest Gemini models from Google, enables users to detect the comments they care about, which they can then label or analyze.

Because we still have a lot to learn, we're inviting developers and researchers to join us in a limited beta program to experiment with these customizable attributes. This collaborative approach will allow us to provide technical support, gather valuable feedback, and iterate on our research in real-world scenarios.

As always, let us know of any questions by contacting us here.

Thank you,