methods, data, analyses | 2026, pp. 1-11 DOI: 10.25521/mda.1065

A Dual-Annotated YouTube Comment Corpus for Studying Sentiment Classification Divergence in Emerging Technology Discourse

Johnny Chan[1], Yilong Wang[1],
Brice Valentin Kok-Shun[1], Meng Wai Woo[1]

[1] *University of Auckland*

Abstract

This data report documents a corpus of 16,652 YouTube comments on precision fermentation technology, each independently classified for sentiment by two automated approaches: GPT-4o, a large language model accessed via prompt engineering, and EmoRoBERTa, a fine-tuned transformer model. The two models disagree on 39.1% of all classifications, and the disagreement is systematically asymmetric: EmoRoBERTa classifies substantially fewer comments as negative than GPT-4o. A subset of 301 comments includes additional human-annotated labels, providing indicative validation of model performance. The corpus spans February 2019 to November 2024, covering the period during which precision fermentation products entered commercial markets. To comply with YouTube API data redistribution policies, the deposited files contain comment identifiers and model-generated labels without comment text, following established rehydration conventions. The full data collection and preprocessing pipeline is documented in a public code repository. This corpus supports research on cross-model annotation reliability, sentiment classification methodology, and computational approaches to studying public discourse on emerging technologies.

Keywords: sentiment analysis, YouTube comments, cross-model disagreement, GPT-4o, EmoRoBERTa

© The Author(s) 2026. This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 License. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Automated sentiment analysis is now widely used in computational social science to study public opinion across digital platforms (Jim et al., 2024; Savci & Das, 2023). In the social sciences, dictionary-based methods that score text against precompiled sentiment lexicons remain a common choice because they are transparent, computationally inexpensive, and require no labeled training data (Hutto & Gilbert, 2014; Taboada et al., 2011). Their main weakness is that they assign sentiment word by word, which leaves them poorly equipped for negation, sarcasm, intensification, and domain-specific usage and causes them to transfer unreliably to contexts whose vocabulary differs from that of the source lexicon (Hartmann et al., 2023; van Atteveldt, van der Velden et al., 2021). Two more recent approaches have gained ground partly in response to these limitations. The first relies on transformer-based models fine-tuned for emotion or sentiment detection, such as variants of RoBERTa trained on labeled emotion corpora (Barbieri et al., 2020). The second, increasingly prevalent, uses general-purpose large language models (LLMs) such as OpenAI's GPT-4o, prompted to perform sentiment classification without task-specific fine-tuning (Krugmann & Hartmann, 2024). Both represent text in context, capturing negation and usage that word-level lexicons miss, and both are now deployed in applied research. Yet systematic comparisons between them on the same data remain limited, particularly in domain-specific contexts where specialized vocabulary, sarcasm, or culturally embedded expressions may challenge general-purpose models.

This gap is consequential. When researchers choose a sentiment classifier, they implicitly assume its outputs are sufficiently reliable for the inferences they wish to draw. If two widely used models produce substantially different classifications on the same data, the choice of model becomes an analytical decision with direct implications for findings. Understanding where and how models diverge is therefore a prerequisite for methodological transparency in computational text analysis.

This data report documents a dual-annotated YouTube comment corpus in which every record has been independently classified by both GPT-4o and EmoRoBERTa. The corpus comprises 16,652 comments from 180 videos on precision fermentation technology, spanning February 2019 to November 2024. The two models disagree on 39.1% of classifications, with the disagreement concentrated in a specific and consequential pattern: EmoRoBERTa systematically under-detects negative sentiment relative to GPT-4o. The dataset is designed to support research on cross-model annotation reliability, sentiment classification methodology, and computational approaches to studying public discourse on emerging technologies.

Direct correspondence to
Johnny Chan, Department of Information Systems and Operations Management, University of Auckland Business School, 12 Grafton Road, Auckland 1010, New Zealand
E-mail: jh.chan@auckland.ac.nz

Data Collection

Source and Scope

Comments were collected via the YouTube Data API (CommentThreads endpoint) from videos related to precision fermentation, a technology that uses engineered microorganisms to produce food ingredients such as dairy proteins. The data collection period spanned January 2018 to November 2024, though the earliest comments in the corpus date from February 2019, reflecting the nascent state of public discussion prior to that point. Video selection followed two complementary procedures. For general coverage of the topic, the YouTube Data API Search endpoint was queried with two terms, one technical ("precision fermentation") and one accessible to general audiences ("lab grown dairy"); each query was restricted to videos, limited to English-language relevance, bounded to the collection window, and ordered by view count, returning up to 50 videos per term. To ensure that discussion of specific commercial products was represented, a further set of product-review videos was identified individually and retrieved by video identifier; these covered branded precision-fermentation products, including Brave Robot ice cream, Perfect Day animal-free dairy, and Modern Kitchen cream cheese. For every selected video, the CommentThreads endpoint was then used to retrieve top-level comments and their replies. Data collection was carried out in Python using an open-source client for the YouTube Data API developed by the research team (available in the code repository), and the retrieved comments were assembled, de-duplicated by comment identifier, and merged using the pandas library.

The resulting corpus comprises 16,652 comments drawn from 180 unique videos. Of these, 12,298 are top-level comments and 4,354 are replies. The videos were created by a range of content producers, including science communicators, industry professionals, food reviewers, and academic commentators. Table 1 provides a summary of the corpus characteristics.

Preprocessing

All comment text was preprocessed prior to sentiment classification. The preprocessing pipeline, documented in the project's code repository, consisted of three steps applied sequentially: (a) lowercasing of all text, (b) removal of special characters and emojis via regular expression filtering (retaining only alphanumeric characters and whitespace), and (c) expansion of contractions using a dictionary-based approach. The pipeline was implemented in Python: Special-character and emoji removal used the standard-library re module for regular-expression filtering, and contraction expansion used a hand-compiled dictionary of contraction forms.

Table 1

Corpus Overview

Attribute	Value
Total comments	16,652
Top-level comments	12,298
Replies	4,354
Unique videos	180
Date range	Feb 2019–Nov 2024
Timestamp precision	Day-level (midnight UTC)

Note. Timestamps are Unix epoch milliseconds but resolve to midnight UTC, indicating day-level granularity.

One preprocessing artifact should be noted. Because special character removal was applied before contraction expansion, apostrophes were stripped before the contraction dictionary was consulted. This means that contractions such as "it's" were first reduced to "its" and then correctly expanded to "it is." However, the contraction dictionary also contained entries for common English words that coincidentally match apostrophe-free contraction forms. For example, "well" was mapped to "we will," "were" to "we are," and "hell" to "he will." These mappings produced false expansions in a substantial number of comments. The word "hell" was expanded to "he will" in approximately 94 comments, "well" to "we will" in approximately 582 comments, and "were" (past tense) to "we are" in approximately 616 comments. Since both sentiment models received the same preprocessed text, this artifact does not differentially bias cross-model comparisons, but it may affect the absolute accuracy of either model's classifications relative to the original comment meaning. Researchers who rehydrate the raw comment text via the YouTube Data API can classify unprocessed text and compare results.

Ethical and Platform Compliance Considerations

All comments in this corpus were publicly posted on a public platform. No personally identifiable information beyond publicly visible display names was collected, and display names are excluded from the deposited files. The research is exempt from institutional ethics review under Section 3.1.1 of the University of Auckland's *Guiding Principles for Conducting Research with Human Participants* (2022), which exempts research using only publicly available data.

To comply with YouTube API Developer Policies regarding data storage and redistribution (Sections III.E.4 and III.G.1), the deposited files contain comment identifiers and analytical labels only, excluding comment text. This follows the

rehydration convention adopted in computational social science research on platform data (Assenmacher et al., 2023; Proferes et al., 2021), in which researchers share unique content identifiers rather than the content itself.

Sentiment Annotation

Automated Annotation

Two models were applied independently to all 16,652 comments. GPT-4o was accessed via the OpenAI API with a prompt-engineered classification task instructing the model to classify each comment as positive, negative, or neutral with respect to precision fermentation and related food technologies. The prompt was designed to constrain outputs to the three permitted categories and to adapt the model to domain-specific language. EmoRoBERTa is a RoBERTa-based model fine-tuned on the GoEmotions dataset (Demszky et al., 2020), accessed through the Hugging Face transformers library, which classifies text into 28 categories (27 emotions plus neutral). These categories were mapped to a three-category scheme: 12 emotions (e.g., joy, admiration, gratitude) were mapped to positive; 11 emotions (e.g., anger, disgust, disappointment) to negative; and 4 ambiguous emotions (e.g., curiosity, surprise) plus neutral. The full mapping and classification code are available in the code repository.

The two models produced markedly different sentiment distributions (Table 2). GPT-4o classified 46.6% of comments as neutral, 34.6% as negative, and 18.9% as positive. EmoRoBERTa classified 60.0% as neutral, 27.2% as positive, and only 12.8% as negative. The models disagreed on the classification of 6,512 comments (39.1%). This disagreement rate is not uniformly distributed across categories: It is concentrated in cases where GPT-4o assigns a negative label and EmoRoBERTa does not.

Table 2

Sentiment Distributions by Annotation Source

Source	*n*	Positive (%)	Neutral (%)	Negative (%)
GPT-4o	16,652	18.9	46.6	34.6
EmoRoBERTa	16,652	27.2	60.0	12.8
Human (subset)	301	27.9	63.1	9.0

Note. Human labels are from a convenience subset of 301 comments drawn from 6 videos (see Section “Human Annotation”).

Human Annotation

A subset of 301 comments was annotated by a single human coder, who classified each comment as positive, negative, or neutral. This subset was drawn from the initial portion of the dataset and covers comments from 6 of the 180 videos, with a single video (a product review of Brave Robot ice cream) contributing 210 of the 301 comments (69.8%). The human-annotated subset is therefore not representative of the full corpus in either video composition or sentiment distribution; it contains proportionally more positive comments and fewer negative comments than the full dataset.

This subset is included in the deposited data because it provides an indicative comparison point for model performance, but it should not be treated as a gold standard. The reliance on a single annotator means that inter-annotator reliability cannot be assessed, and the concentration in a small number of videos limits the generalizability of any accuracy figures derived from it. The human labels are best used for relative comparison between models on the same comments rather than for making generalized claims about either model's absolute accuracy.

Cross-Model Divergence Analysis

Overall Disagreement

The primary analytical contribution of this corpus is the systematic divergence between the two automated classifiers. Across all 16,652 comments, GPT-4o and EmoRoBERTa agreed on 10,140 classifications (60.9%) and disagreed on 6,512 (39.1%). This disagreement rate is substantial and indicates that the choice of sentiment classifier is not methodologically neutral: It can materially affect the findings of any downstream analysis.

Asymmetric Negative Sentiment Detection

The disagreement between models is not uniformly distributed across sentiment categories. The most consequential divergence occurs in negative sentiment detection. GPT-4o classified 5,756 comments (34.6%) as negative, while EmoRoBERTa classified only 2,136 (12.8%), a difference of 3,620 comments. The indicative validation from the human-annotated subset supports the interpretation that GPT-4o is closer to human judgment on negative sentiment: Among the 27 comments the human coder labeled negative, GPT-4o correctly identified 26 (96.3%), while EmoRoBERTa identified only 11 (40.7%). Table 3 presents the confusion matrices for both models against human labels.

Table 3

Confusion Matrices: GPT-4o and EmoRoBERTa Against Human Labels (n = 301)

Panel A: GPT-4o (overall accuracy: 75.7%)

	Predicted (GPT-4o)			
Actual (Human)	Positive	Neutral	Negative	Accuracy
Positive (*n* = 84)	69	14	1	82.1%
Neutral (*n* = 190)	33	133	24	70.0%
Negative (*n* = 27)	1	0	26	96.3%

Panel B: EmoRoBERTa (overall accuracy: 66.4%)

	Predicted (EmoRoBERTa)			
Actual (Human)	Positive	Neutral	Negative	Accuracy
Positive (*n* = 84)	66	16	2	78.6%
Neutral (*n* = 190)	56	123	11	64.7%
Negative (*n* = 27)	4	12	11	40.7%

Note. Human labels are from a convenience subset (6 videos, single annotator). Per-class accuracy figures should be interpreted with caution given the small number of negative cases (*n* = 27).

This pattern has substantive implications for applied research. In domains where public skepticism, criticism, and opposition are of analytical interest (as they are in discourse surrounding emerging food technologies, biotechnology, and sustainability), a classifier that systematically undercounts negative sentiment will produce a materially distorted picture of public opinion. Of the 27 comments the human coder labeled as negative, EmoRoBERTa misclassified 12 as neutral and 4 as positive, rendering nearly 60% of negative sentiment invisible.

Agreement and Ensemble Potential

When the two models agreed on a classification (10,140 of 16,652 comments), the indicative validation suggests their joint accuracy against human labels was 81.3% (174 of 214 agreeing cases in the human-annotated subset). This finding, while based on the non-representative convenience subset, points toward the potential value of ensemble approaches in which classifications are accepted only when multiple models converge. The 6,512 comments on which the models disagree represent a particularly interesting subset for further research, as they

are likely to contain nuanced, ambiguous, or mixed sentiment expressions that challenge automated classification.

Dataset Description and Access

Deposited Files

The corpus is deposited as two separate files reflecting distinct research purposes.

File 1: Human-annotated subset (n = 301). This file contains the three-way annotated subset. Fields: VideoID, CommentID, Timestamp, GPT4o_Label, EmoRoBERTa_Label, Human_Label. Each label takes one of three values: positive, negative, or neutral.

File 2: Full corpus (n = 16,652). This file contains all collected comments with dual model annotations. Fields: VideoID, CommentID, Timestamp, Reply_Flag, GPT4o_Label, EmoRoBERTa_Label.

Comment text is excluded from both files to comply with YouTube's data redistribution policies. All sentiment labels are normalized to lowercase (positive, negative, neutral). Timestamps are Unix epoch milliseconds at day-level granularity.

Rehydration

Researchers wishing to obtain the original comment text may query the YouTube Data API's Comments endpoint using the provided CommentIDs. Some comments may have been deleted or made private since the original collection in 2024; this attrition is an inherent characteristic of the rehydration approach (Assenmacher et al., 2023). Four VideoIDs in the dataset begin with a hyphen character; researchers should take care that these are not corrupted by spreadsheet software interpreting them as formula expressions.

Rehydration is not the only route to working with data whose text cannot be redistributed. Where terms of service or copyright prevent sharing the underlying content, arrangements such as non-consumptive use, in which researchers compute over text they cannot read or export, and remote code execution, in which analysis code is sent to run against the data in a controlled environment, allow reproducible analysis without transferring the protected text itself (van Atteveldt, Althaus, et al., 2021). These arrangements demand more infrastructure and coordination than sharing identifiers, but they preserve access to the original text and avoid the attrition that rehydration inevitably introduces. We

note them here as complementary directions for future deposits of platform data of this kind.

Potential Applications

The corpus supports several research applications. First, the dual annotations across the full corpus enable systematic study of where and how different sentiment classification approaches diverge, without requiring human labels. Researchers can examine disagreement patterns by comment length, reply status, video context, or temporal period. Second, the model disagreement subset (6,512 comments) constitutes a resource for studying ambiguous or challenging cases in automated text annotation, supporting the development of improved classifiers or ensemble strategies. Third, the temporal metadata spanning nearly six years enables longitudinal analysis of how automated sentiment classifications of public discourse on an emerging technology evolve as the technology moves from experimental to commercial stages.

Limitations

Several limitations should be considered. The preprocessing pipeline introduced false contraction expansions affecting a portion of comments, as documented in Section “Preprocessing”; researchers can mitigate this by rehydrating and reclassifying raw text. A related limitation concerns emojis, which were removed during preprocessing along with other special characters. Emojis frequently carry, intensify, or even invert the sentiment of a comment, so their removal may have suppressed affective signals, particularly for ironic or negative expressions. The deposited labels therefore reflect the cleaned text and may diverge from the sentiment a reader would infer from the original comment. This limitation is more tractable than it once was, because recent large language models interpret emojis directly. Researchers who rehydrate the raw comments can classify the unprocessed text with emojis retained, which may recover sentiment that the present pipeline discards. The human-annotated subset is a convenience sample from 6 of 180 videos with a single annotator, limiting both its representativeness and its utility as a gold-standard reference. The corpus is restricted to English-language comments, excluding non-English discourse. The rehydration approach means the recoverable text content will degrade over time. Precision fermentation is a single technology domain, and the cross-model divergence patterns observed here may not generalize to other subject areas. Finally, both models classified preprocessed rather than raw text, meaning the deposited labels reflect model performance on cleaned input. Results on unprocessed text may differ.

Data and Code Availability

Both datasets (human-annotated subset and full corpus) are deposited at Zenodo, accessible at https://zenodo.org/records/21146055. The complete data collection, preprocessing, and sentiment classification pipeline is publicly available at https://github.com/isom-ds/ffsp-sm-emotions. Comment text is excluded from the deposited files in accordance with YouTube API Services Developer Policies; researchers may rehydrate the text using the provided CommentIDs via the YouTube Data API. A plain-text codebook describing every variable in both deposited files is included with the Zenodo deposit.

Conflict of Interest

The authors declare no conflicts of interest.

Ethics

This study did not involve direct interaction with human participants and used only publicly available YouTube comments; the deposited files contain no comment text or personally identifiable information. The research is exempt from ethics review under the University of Auckland's guiding principles for research with human participants, which exempt research using only publicly available data.

Funding

This work was supported by the Future Food Research Centre, University of Auckland, New Zealand [Seed Funding 2024].

Use of AI Tools

GPT-4o (OpenAI) was used as one of two automated sentiment classifiers applied to the corpus, as described in Section "Automated Annotation". Claude (Anthropic) was used to assist in proofreading and polishing the manuscript. The authors reviewed and edited all content and take full responsibility for the publication.

References

Assenmacher, D., Sen, I., Fröhling, L., & Wagner, C. (2023). The end of the rehydration era—The problem of sharing harmful Twitter research data. *Workshop Proceedings of the 17th International AAAI Conference on Web and Social Media, 2023*, 56. https://doi.org/10.36190/2023.56

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1644–1650). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2023). More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, *40*(1), 75–87. https://doi.org/10.1016/j.ijresmar.2022.05.005

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, *8*(1), 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, *6*, Article 100059. https://doi.org/10.1016/j.nlp.2024.100059

Krugmann, J. O., & Hartmann, J. (2024). Sentiment analysis in the age of generative AI. *Customer Needs and Solutions*, *11*(1), Article 3. https://doi.org/10.1007/s40547-024-00143-4

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, *7*(2), Article 20563051211019004. https://doi.org/10.1177/20563051211019004

Savci, P., & Das, B. (2023). Prediction of the customers' interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages. *Journal of King Saud University - Computer and Information Sciences*, *35*(3), 227–237. https://doi.org/10.1016/j.jksuci.2023.02.017

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

University of Auckland. (2022). *Guiding Principles for Conducting Research with Human Participants*.

van Atteveldt, W., Althaus, S., & Wessler, H. (2021). The trouble with sharing your privates: Pursuing ethical open science and collaborative research across national jurisdictions using sensitive data. *Political Communication*, *38*(1–2), 192–198. https://doi.org/10.1080/10584609.2020.1744780

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

Online First publication date: 2026-11-03