Investigating Indonesian Public Discourse in YouTube-Based Science Communication: A Corpus-Based Computational Linguistics Approach

Muhamad Farrel Dava Fauzan*
[1]*Physics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta*

ABSTRACT
Digital platforms have expanded science communication by enabling audiences to interact with scientific content. However, computational resources for examining Indonesian-language audience discourse remain limited. This study presents the Indonesian Public Discourse Corpus (IPDC), constructed from 202,429 YouTube comments from 35 science education videos on the Kok Bisa? channel. The research employed a multi-stage natural language processing workflow involving corpus acquisition, manual annotation, transformer-based classification, and full-corpus inference. A dataset of 10,500 comments was independently annotated into eight discourse-act categories: Question, Opinion, Disagreement, Correction, Suggestion, Praise, Agreement, and Experience. The annotation protocol calibration achieved 93.00% agreement (Cohen's Kappa = 0.9122). The strongest model, IndoBERT Base, achieved a Macro F1 of 97.40%, a Weighted F1 of 97.72%, and an accuracy of 97.73% on unseen Kok Bisa? videos. Full-corpus inference showed Opinion was the dominant discourse act (69.20%), followed by Question (22.59%). In contrast, Experience, representing only 2.18% of comments, received the highest mean number of likes. The findings demonstrate the potential of transformer-based classification for large-scale analysis of audience participation in Indonesian digital science communication.

ABSTRAK
Platform digital telah memperluas komunikasi sains dengan memungkinkan audiens berinteraksi dengan konten ilmiah. Namun, sumber daya komputasional untuk mengkaji wacana audiens berbahasa Indonesia masih terbatas. Penelitian ini menyajikan *Indonesian Public Discourse Corpus* (IPDC), yang dibangun dari 202.429 komentar YouTube pada 35 video edukasi sains di kanal Kok Bisa?. Penelitian menggunakan alur kerja *natural language processing* yang mencakup akuisisi korpus, anotasi manual, klasifikasi berbasis *transformer*, dan inferensi pada seluruh korpus. Sebanyak 10.500 komentar dianotasi secara independen menjadi delapan kategori wacana: *Question, Opinion, Disagreement, Correction, Suggestion, Praise, Agreement*, dan *Experience*. Kalibrasi protokol anotasi menghasilkan tingkat kesepakatan 93,00% (Cohen's Kappa = 0,9122). Model terbaik, IndoBERT Base, mencapai Macro F1 97,40%, Weighted F1 97,72%, dan akurasi 97,73% pada data uji dari video *Kok Bisa?* yang belum pernah dilihat model. Inferensi seluruh korpus menunjukkan *Opinion* sebagai wacana paling dominan (69,20%), diikuti *Question* (22,59%). Sebaliknya, kategori *Experience* (2,18%) memperoleh rata-rata jumlah *likes* tertinggi. Temuan ini membuktikan potensi klasifikasi berbasis *transformer* untuk menganalisis partisipasi audiens berskala besar dalam komunikasi sains digital di Indonesia.

CONTACT
mfarreldava23@gmail.com

KEYWORDS
Computational Linguistics, Digital Science Communication, Discourse Analysis, *IndoBERT*, YouTube

INTRODUCTION

Digital media have expanded the contexts in which scientific information is communicated and discussed. Contemporary science communication therefore extends beyond conventional one-way dissemination and increasingly involves relationships among scientific institutions, communicators, media, and different publics (Bucchi & Trench, 2021). Digital platforms are particularly relevant to this transformation because they allow

audiences not only to access scientific content but also to respond publicly through comments and other forms of interaction.

YouTube has become an important environment for science communication and public engagement. Research on science communication on YouTube has shown that the platform supports substantial public interaction with science content and that emotional and cognitive engagement may both be associated with observable forms of participation, including commenting (Dubovi & Tabak, 2021a). Research examining YouTube comments has further demonstrated that comment sections can provide opportunities for discussion and knowledge co-construction rather than functioning solely as spaces for isolated reactions (Dubovi & Tabak, 2020). From the perspective of science communicators themselves, comments are also regarded as an important component of audience interaction and perceived communication impact (Hill et al., 2022).

Within the Indonesian digital media environment, Kok Bisa? provides a relevant case for examining audience participation around publicly accessible science education content. The channel publishes videos that communicate scientific and educational topics to a broad Indonesian-speaking audience and provides a large-scale public comment environment in which viewers can express questions, opinions, disagreements, corrections, suggestions, praise, agreement, and personal experiences. The comment environment associated with Kok Bisa? therefore provides an opportunity to examine how Indonesian-language audiences participate in science-related public discourse at corpus scale.

The increasing scale of digital audience interaction creates a methodological challenge for research. Manual analysis can provide detailed interpretation of individual comments, but it becomes increasingly difficult to apply when datasets contain tens or hundreds of thousands of observations. Computational approaches can extend textual analysis to larger collections. Transformer-based language models have demonstrated the ability to learn contextual representations that can subsequently be adapted to specific natural language processing tasks (Devlin et al., 2019).

However, approaches such as sentiment classification do not necessarily identify what a participant is doing communicatively. A positive comment may express praise, agreement, or a personal experience, whereas a negative comment may represent disagreement or an attempt to correct a factual claim. Distinguishing these functions requires attention to the communicative act performed through language. This perspective is grounded in speech-act theory, which conceptualizes utterances as actions that can perform different communicative functions rather than merely convey propositional content (Austin, 1962; Searle, 1969). Computational research has also demonstrated the feasibility of classifying speech acts in large-scale user-generated discourse (Vosoughi & Roy, 2016)

The Indonesian context presents additional challenges for computational discourse analysis. Indonesia is characterized by substantial linguistic diversity, while Indonesian natural language processing continues to face limitations related to resources and linguistic variation (Aji et al., 2022). User-generated Indonesian text may also contain informal spelling, non-standard lexical forms, and code-mixing, which can complicate text normalization and automated language processing (Barik et al., 2019). These characteristics make contextual language models particularly relevant to the analysis of Indonesian-language digital discourse.

Indonesian pre-trained language models have been developed to address this need. IndoBERT was introduced as part of the IndoLEM framework and provides contextual representations specifically designed for Indonesian natural language processing tasks (Koto et al., 2020). Such models create an opportunity to move beyond surface-level keyword matching and develop classification systems capable of distinguishing context-dependent discourse functions.

Although previous studies have examined science communication on YouTube, public engagement with science, and Indonesian natural language processing, these areas have rarely been integrated to examine the communicative functions of Indonesian-language audience participation at corpus scale (Aji et al., 2022; Dubovi & Tabak, 2020, 2021a; Hill et al., 2022; Koto et al., 2020). Moreover, large-scale discourse analysis remains limited in Indonesian-language science communication contexts.

The present study addresses this intersection by developing the Indonesian Public Discourse Corpus (IPDC) from publicly accessible YouTube comments posted on science education videos published by Kok Bisa?. The study develops an eight-category discourse taxonomy, constructs a manually annotated gold-standard dataset, evaluates transformer-based classification models using a video-stratified design, and applies the selected model to the complete corpus. Rather than claiming to represent all Indonesian science communication, the study uses the Kok Bisa? comment environment as a large-scale empirical case through which computational discourse analysis can be developed and evaluated.

METHODS

This study employed a multi-stage computational workflow consisting of corpus construction, screening and preprocessing, discourse taxonomy development, manual annotation, video-stratified dataset partitioning, transformer fine-tuning, model benchmarking, and full-corpus inference.

The corpus was constructed from publicly accessible comments associated with science education videos published on the Kok Bisa? YouTube channel. An initial collection identified 36 videos within the selected source collection. One video was unavailable during the data acquisition process and was excluded from the analysis. The final corpus therefore consisted of comments associated with 35 valid publicly accessible videos.

The final dataset contained 202,429 comments, consisting of 116,459 top-level comments and 85,970 threaded replies. The research artifact also recorded 4,408 synchronized transcript segments associated with the analyzed Kok Bisa? video collection.

Corpus preparation included text cleaning, normalization, spam detection, duplicate identification, and language screening. A total of 41 language codes were detected, with Indonesian accounting for 159,281 comments, or 78.68% of the corpus. Spam filtering identified 2,574 promotional or bot-like comments, corresponding to 1.27% of the collected data, while 12,380 exact duplicate texts, or 6.12%, were identified during corpus preparation. The final cleaned corpus contained 3,059,944 tokens.

The preprocessing procedure was designed to reduce textual noise while retaining information potentially relevant to discourse interpretation. The workflow included regular-expression-based cleaning and normalization of informal lexical forms using a 1,247-entry slang lexicon. Emoji were retained because they may contribute to the pragmatic meaning of user-generated comments. The presence of informal language, non-standard forms, and code-mixing is a recognized challenge in processing Indonesian user-generated text (Barik et al., 2019). Accordingly, automated language identification was treated as a screening component rather than as the sole determinant of whether a comment could contribute to discourse analysis.

The discourse taxonomy was conceptually informed by speech-act theory, which distinguishes communicative actions performed through language (Austin, 1962; Searle, 1969). Each comment in the gold-standard dataset was assigned one primary discourse-act category: *Question, Opinion, Disagreement, Correction, Suggestion, Praise, Agreement,* or *Experience*. The final annotated dataset consisted of 10,500 comments.

Two annotators independently participated in the annotation process. Inter-annotator reliability was assessed using Cohen's Kappa, a coefficient designed to evaluate agreement between raters for nominal classifications while accounting for agreement expected by chance (Cohen, 1960). An initial pilot annotation of 520 comments produced a raw agreement of 44.23% and Cohen's Kappa of $\kappa = 0.3027$. This result was not treated as the final reliability estimate. Instead, disagreement patterns were examined to identify recurring ambiguity in category boundaries.

The annotation protocol was subsequently revised into Codebook v1.1. The revised protocol prioritized *illocutionary force over surface syntax*, meaning that comments were categorized according to their primary communicative function rather than punctuation or isolated lexical forms. Three recurrent boundary problems were explicitly addressed. First, rhetorical questions expressing a challenge or assertion were classified according to their underlying communicative function. Second, *Correction* required a substantive factual, numerical, formulaic, or other corrective proposition, whereas unsupported opposition was classified as *Disagreement*. Third, comments requesting or proposing future content or action were classified as *Suggestion* rather than *Question*.

A subsequent calibration procedure involving 100 difficult boundary cases produced 93.00% agreement and Cohen's Kappa of $\kappa = 0.9122$. During the final gold-standard annotation process, agreement estimates ranged from $\kappa = 0.88$ to $\kappa = 0.91$.

The final gold-standard dataset contained 4,007 *Opinion* comments (38.2%), 1,246 *Question* comments (11.9%), 994 *Disagreement* comments (9.5%), 968 *Praise* comments (9.2%), 961 *Suggestion* comments (9.2%), 939 *Experience* comments (8.9%), 714 *Correction* comments (6.8%), and 671 *Agreement* comments (6.4%). The maximum-to-minimum class ratio was 5.97. Class-weighted cross-entropy loss was therefore used during model training to reduce the influence of class imbalance.

To reduce contextual overlap between training and evaluation, comments were partitioned using a video-stratified group split. Comments originating from the same Kok Bisa? video were assigned to only one dataset partition. The training set contained 7,148 comments from 24 videos, the validation set contained 1,850 comments from 5 videos, and the held-out test set contained 1,502 comments from the remaining 6 videos.

The primary transformer model was IndoBERT Base Phase 1, which is part of the Indonesian pre-trained language model ecosystem introduced by Koto et al. (2020). The experimental comparison also included two fine-tuned mDeBERTa-v3 Base configurations and a zero-shot mDeBERTa-v3 XNLI reference.

The IndoBERT experiments evaluated learning rates of 1×10^{-5}, 2×10^{-5}, and 3×10^{-5}. Training used the AdamW optimizer, a maximum sequence length of 128 tokens, weight decay of 0.01, a warmup ratio of 0.10, cosine learning-rate decay with linear warmup, and early stopping based on validation Macro F1 with a patience of three epochs. The IndoBERT experiments used a batch size of 16, whereas the mDeBERTa baseline experiments used a batch size of 8.

Model performance was evaluated using Macro F1, Weighted F1, and accuracy. Macro F1 was included to assess performance across categories without allowing the majority class to dominate the evaluation. Weighted F1 complemented this metric by accounting for class frequency. The strongest-performing model was subsequently applied to all 202,429 comments collected from the Kok Bisa? video corpus to estimate the distribution of discourse acts. Predicted discourse categories were also compared descriptively with mean comment likes as an observable engagement indicator.

RESULTS AND DISCUSSIONS

The results are presented in a sequence that moves from corpus characterization and model evaluation to full-corpus inference and cross-video comparison. The findings therefore distinguish between the linguistic properties of the Kok Bisa? comment corpus, the performance of the classification system, and the substantive patterns of audience discourse identified after large-scale inference.

Before evaluating the classification results, the lexical structure of the IPDC was examined through its empirical rank-frequency distribution. Figure 1 shows the distribution of token frequencies across 3,059,944 tokens on logarithmic axes. The observed distribution exhibited a pronounced heavy-tailed pattern: a relatively small number of highly frequent words occupied the head of the distribution, whereas a substantially larger number of lexical items occurred at progressively lower frequencies.

The fitted relationship followed a power-law-like pattern, expressed as $f(r) \propto r^{-1.3171}$, with an R^2value of 0.98. The lower residual panel indicates that the fitted relationship provides a close descriptive approximation across much of the observed rank range, although deviations remain at the extremes of the distribution. These results characterize the Kok Bisa? comment corpus as having a highly uneven lexical frequency structure, which is relevant to the computational analysis of large-scale natural language data.

The long-tail region should not be interpreted as consisting exclusively of technical physics or science vocabulary. Rather, it contains lower-frequency lexical items that may include domain-specific expressions alongside less frequent general-language forms. Therefore, the distribution is presented as a descriptive property of the corpus rather than as evidence that every low-frequency token represents scientific terminology.

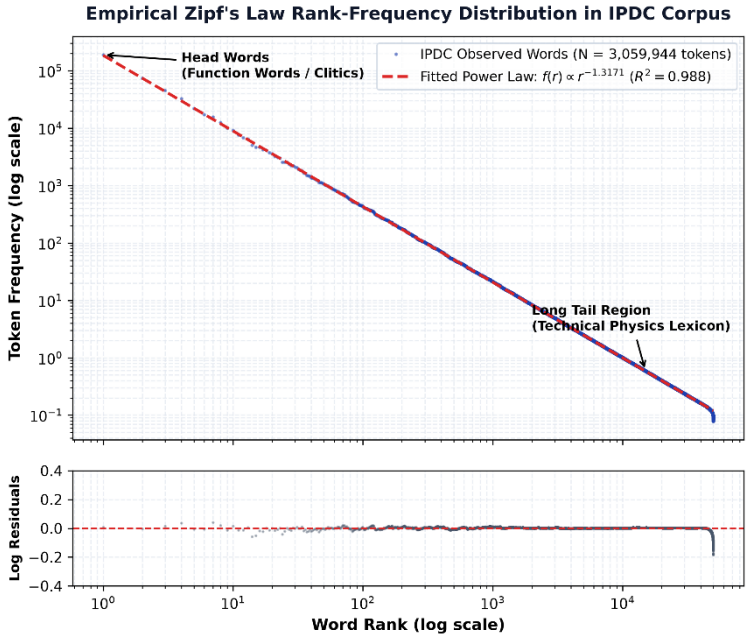


Figure 1. Empirical rank-frequency distribution of tokens in the Indonesian Public Discourse Corpus (IPDC), constructed from comments on Kok Bisa? YouTube videos.

The model comparison was conducted on the held-out test partition containing 1,502 comments associated with six Kok Bisa? videos that were not used during training or validation.

Table 1. Performance of the Evaluated Models on the Held-Out Test Set

Model	Learning Rate	Macro F1 (%)	Weighted F1 (%)	Accuracy (%)
Zero-Shot mDeBERTa-v3 XNLI	—	42.21	44.97	43.65
mDeBERTa-v3 Base Baseline A	1×10^{-5}	96.85	97.21	97.25
mDeBERTa-v3 Base Baseline B	2×10^{-5}	97.12	97.50	97.54
IndoBERT Base	1×10^{-5}	95.82	96.54	96.60
IndoBERT Base	2×10^{-5}	96.93	97.39	97.41
IndoBERT Base	3×10^{-5}	97.40	97.72	97.73

The strongest configuration was IndoBERT Base with a learning rate of 3×10^{-5}, achieving a Macro F1 of 97.40%, a Weighted F1 of 97.72%, and an accuracy of 97.73%. This configuration produced an observed Macro F1 difference of 0.28 percentage points relative to the strongest evaluated mDeBERTa-v3 baseline. Because no statistical significance test was conducted, the difference is reported as an observed performance difference rather than as statistically significant superiority.

The gap between the zero-shot reference and the supervised models was substantially larger. The zero-shot configuration obtained a Macro F1 of 42.21%, whereas the champion IndoBERT model obtained 97.40%, an observed difference of 55.19 percentage points. Within the present evaluation, this pattern indicates that supervised task-specific training was considerably more effective than the evaluated zero-shot configuration for distinguishing the eight discourse categories in the Kok Bisa? comment dataset.

The results are consistent with the broader rationale for adapting contextual language models to specific downstream tasks. BERT-based models are designed to learn bidirectional contextual representations that can subsequently be fine-tuned for task-specific classification (Devlin et al., 2019). In the present study, the use of an Indonesian-specific pre-trained model was also methodologically appropriate because IndoBERT was developed specifically to support Indonesian language understanding tasks (Koto et al., 2020). The training dynamics of the selected IndoBERT Base configuration are presented in Figure 2. Training Macro F1 increased progressively from approximately 76.2% during the first epoch to approximately 98.5% at the final epoch. Validation Macro F1 followed a similar trajectory, increasing from approximately 74.8% to a peak of approximately 97.4%.

The held-out test Macro F1 of 97.40% provides the final performance estimate on comments originating from Kok Bisa? videos that were not included in the training or validation partitions. The convergence of the training and validation trajectories, together with the comparable held-out test performance, indicates that the selected configuration achieved strong performance within the video-stratified evaluation design. However, the test result should be interpreted as evidence of generalization to previously unseen Kok Bisa? videos within the present corpus rather than as external validation across independently collected datasets.

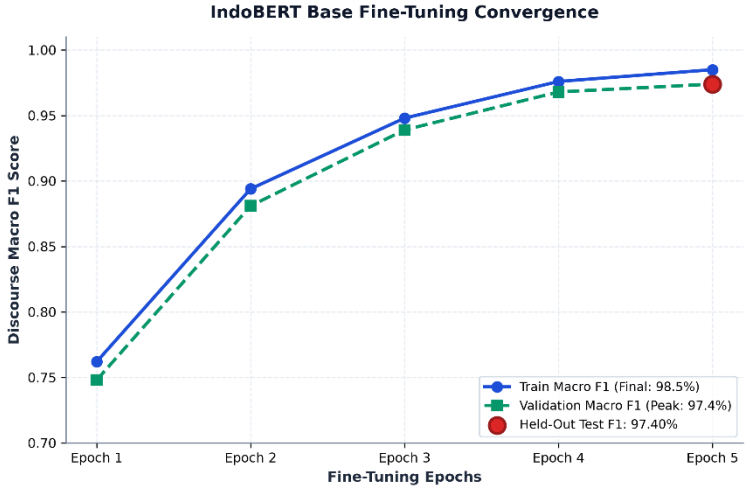


Figure 2. Fine-tuning convergence of the champion IndoBERT Base configuration across five training epochs, evaluated using the Kok Bisa? video-stratified corpus design.

Class-level performance remained high across all eight discourse categories.

Table 2. Class-Level Performance of the Champion IndoBERT Model

Discourse Act	Precision (%)	Recall (%)	F1 (%)
Question	98.2	98.6	98.4
Praise	98.1	97.9	98.0
Suggestion	97.8	97.5	97.6
Disagreement	96.8	97.6	97.2
Experience	97.2	96.3	96.7
Agreement	96.7	96.0	96.3
Correction	95.9	96.5	96.2
Opinion	96.4	95.8	96.1

The high performance of *Correction* and *Disagreement* is particularly relevant because distinguishing these categories was one of the recurring boundary problems identified during pilot annotation. Codebook v1.1 explicitly defined *Correction* as requiring a substantive corrective proposition, whereas opposition without such a proposition was categorized as *Disagreement*. The resulting held-out performance suggests that the revised operational definitions provided sufficiently distinct category boundaries for the model within the present Kok Bisa? dataset.

After model selection, the champion IndoBERT configuration was applied to the complete corpus of 202,429 comments.

Table 3. Full-Corpus Distribution of Predicted Discourse Acts and Mean Likes

Discourse Act	Comments	Percentage (%)	Mean Likes
Opinion	140,092	69.20	2.38
Question	45,738	22.59	2.82
Praise	4,484	2.22	5.24
Experience	4,415	2.18	8.15
Suggestion	3,090	1.53	6.65
Disagreement	2,559	1.26	1.63
Agreement	1,177	0.58	2.94
Correction	874	0.43	2.91
Total	**202,429**	**100.00**	—

The full-corpus analysis showed a highly concentrated distribution. *Opinion* was the dominant predicted discourse act, representing 140,092 comments or 69.20% of the Kok Bisa? corpus. *Question* was the second-largest category, accounting for 45,738 comments or 22.59%. Together, these two categories represented 91.79% of all analyzed comments.

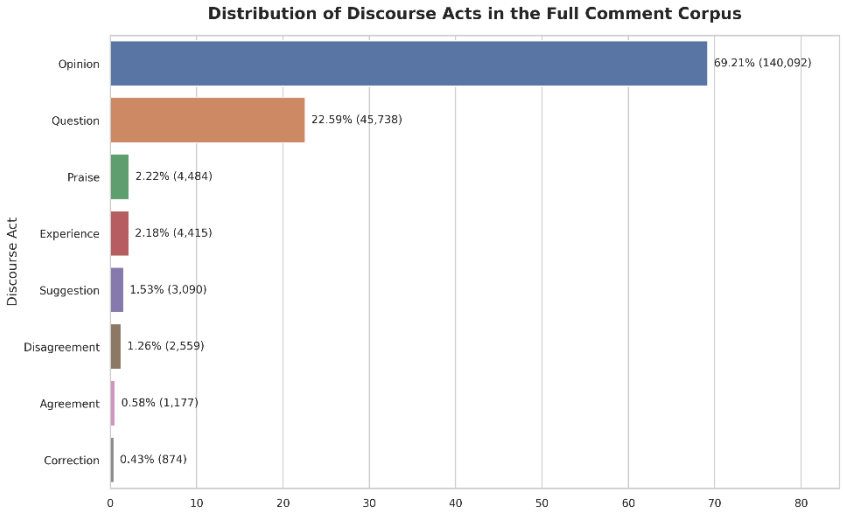


Figure 3. Distribution of predicted discourse acts across the full Indonesian Public Discourse Corpus constructed from Kok Bisa? YouTube comments (N = 202,429).

Figure 3 provides a visual representation of this concentration across the complete Kok Bisa? comment corpus. The distribution shows that audience participation in this analyzed comment environment was primarily expressed through viewpoint articulation and information seeking, while the remaining discourse acts individually represented relatively small proportions of the classified comments.

This result indicates that participation in the Kok Bisa? comment environment was primarily expressed through viewpoint articulation and information seeking. The finding is relevant to previous work describing YouTube as an environment in which audiences actively engage with science rather than functioning only as passive recipients of audiovisual information (Dubovi & Tabak, 2020, 2021a). However, the present classification framework identifies communicative function rather than scientific quality. Therefore, the predominance of *Opinion* does not indicate that these comments were necessarily scientifically accurate, and the prevalence of *Question* does not independently measure scientific literacy or curiosity.

The distribution also demonstrates why discourse analysis can provide information that would be obscured by simple sentiment classification. An *Opinion* may be positive, negative, or neutral in affective orientation, but its primary function within the present taxonomy is the expression of a viewpoint. Similarly, a *Question* is defined by information-seeking or interrogative communicative intent rather than emotional valence. This distinction is consistent with the speech-act perspective that communicative meaning cannot be reduced solely to the propositional or affective properties of an utterance (Austin, 1962; Searle, 1969).

The relationship between discourse frequency and observable engagement showed a different pattern. Although *Experience* represented only 2.18% of the corpus, comments classified into this category received the highest mean number of likes at 8.15. *Suggestion* followed with a mean of 6.65 likes, while *Praise* received 5.24 likes. In contrast, *Opinion*, despite accounting for more than two-thirds of all Kok Bisa? comments, received a mean of 2.38 likes.

The descriptive pattern indicates that the most frequent form of participation was not necessarily associated with the highest observed engagement. This finding is relevant to research showing that engagement with science content on YouTube may involve multiple behavioral, emotional, and cognitive dimensions rather than a single uniform form of interaction (Dubovi & Tabak, 2021a). Nevertheless, the present results should not be interpreted causally. Comment likes may be influenced by factors not modeled in this study, including video popularity, comment age, comment visibility, thread position, and platform-specific exposure dynamics. The mean-like comparison therefore describes an observed engagement pattern rather than demonstrating audience preference or causal influence.

The corpus-level distribution provides an aggregate representation of discourse within the Kok Bisa? comment environment, but aggregate proportions may conceal variation between individual video contexts. To examine whether the overall pattern was broadly maintained across highly active discussion environments, the discourse composition of the seven most-discussed Kok Bisa? videos was compared proportionally.

Figure 4 shows that *Opinion* and *Question* remained the dominant discourse functions across the seven most-discussed videos, although their relative proportions varied substantially between individual Kok Bisa? videos. Some videos displayed a particularly strong concentration of *Opinion*, whereas others contained comparatively larger proportions of *Question* or other discourse acts.

This variation indicates that the corpus-level distribution should not be interpreted as a uniform pattern reproduced identically across all videos published by Kok Bisa?. Instead, the communicative composition of audience participation appears to vary according to the specific discussion environment associated with individual videos. Differences in topic, framing, audience composition, comment visibility, and interaction dynamics may all contribute to this variation, although the present study does not isolate the causal contribution of these factors.

The comparison is consistent with the broader view that public engagement with science on digital platforms can involve different forms of cognitive, emotional, and communicative participation rather than a single homogeneous mode of interaction (Dubovi & Tabak, 2020, 2021a). Within the present study, however, Figure 4 should be interpreted descriptively: it demonstrates variation in the distribution of predicted discourse acts across highly discussed Kok Bisa? videos rather than establishing causal relationships between video characteristics and discourse composition.

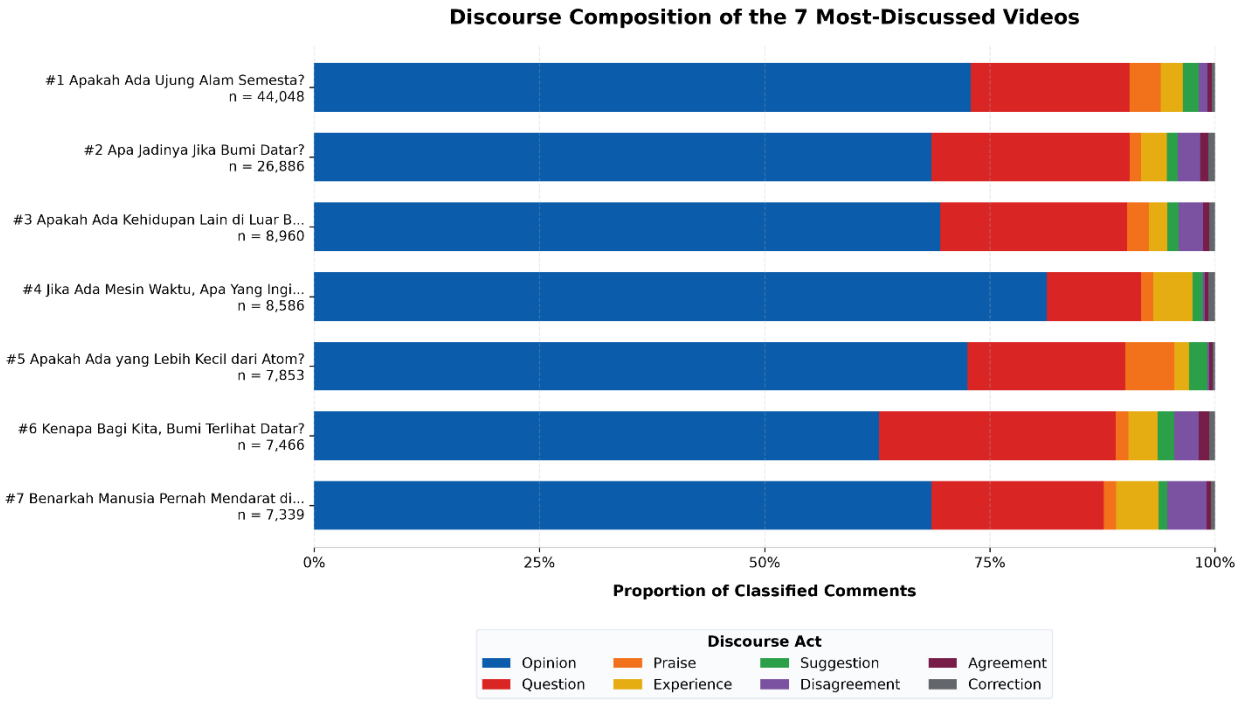


Figure 4. Composition of predicted discourse acts across the seven most-discussed Kok Bisa? videos. Each horizontal bar represents the proportional distribution of discourse categories within an individual video.

The study also demonstrates the value of combining manual annotation with large-scale computational inference. Manual annotation established the operational basis of the discourse categories, whereas the trained model enabled analysis to extend from 10,500 labeled comments to the complete corpus of 202,429 comments posted within the selected Kok Bisa? video environment. The video-stratified partition further reduced overlap between source-video contexts used for training and those used for evaluation.

Several limitations should be considered. First, the corpus represents audience discourse associated specifically with the selected Kok Bisa? science education videos and does not establish that the same distribution will occur across other Indonesian science communicators, scientific domains, or digital platforms. Second, the held-out evaluation involved previously unseen videos within the same broader Kok Bisa? corpus and therefore does not constitute external validation using an independently collected dataset. Third, discourse-act classification does not independently determine factual correctness, misinformation, scientific literacy, or conceptual understanding. Finally, likes were used only as a descriptive observable engagement indicator and should not be interpreted as a direct measure of audience approval.

Future research could apply the framework to additional Indonesian science communicators, compare discourse patterns across scientific domains and communication styles, examine longitudinal changes in audience participation, and evaluate performance on independently collected external corpora. Another important direction would be to combine discourse classification with expert assessment of scientific accuracy or conceptual quality, enabling future work to distinguish not only what audiences are communicatively doing but also the epistemic characteristics of what they are communicating.

To enhance research transparency and provide access to supplementary analytical materials, an interactive digital research companion was developed alongside this study. The companion documents the IPDC corpus, data-processing workflow, annotation framework, model evaluation, and additional visualizations that complement the analyses reported in this article. The accompanying research resource is available at Kok Bisa? Physics Discourse Analytics (Fauzan, 2026)

CONCLUSIONS

This study constructed the Indonesian Public Discourse Corpus (IPDC) from 202,429 publicly accessible YouTube comments associated with 35 science education videos published on the Kok Bisa? YouTube channel. The corpus contained 3,059,944 cleaned tokens and exhibited a strongly uneven rank-frequency structure characterized by a power-law-like relationship between token rank and frequency.

Audience comments were operationalized into eight primary discourse-act categories: *Question, Opinion, Disagreement, Correction, Suggestion, Praise, Agreement,* and *Experience*. A gold-standard dataset of 10,500 comments was independently annotated by two annotators. The annotation protocol was refined through pilot testing, systematic disagreement analysis, and the development of Codebook v1.1. The initial pilot stage identified substantial ambiguity in category boundaries, whereas the revised calibration procedure achieved 93.00% agreement and Cohen's Kappa of $\kappa = 0.9122$.

The strongest evaluated model, IndoBERT Base, achieved a Macro F1 of 97.40%, a Weighted F1 of 97.72%, and an accuracy of 97.73% when evaluated on comments associated with previously unseen Kok Bisa? videos. The training and validation trajectories also showed progressive convergence during the five-epoch fine-tuning process.

Full-corpus inference showed that *Opinion* was the dominant predicted discourse act at 69.20%, followed by *Question* at 22.59%. Together, these categories accounted for 91.79% of the complete Kok Bisa? comment corpus. However, the cross-video comparison demonstrated that the proportional composition of discourse was not identical across individual high-discussion videos. The distribution of observable engagement also differed from discourse frequency, as *Experience* represented only 2.18% of comments but received the highest mean number of likes.

Overall, the findings demonstrate that transformer-based discourse classification can support large-scale analysis of audience participation within Indonesian digital science communication. More specifically, the study establishes Kok Bisa? as the empirical case through which the IPDC was constructed and the proposed computational framework was evaluated. The resulting framework provides both a method for identifying communicative functions at corpus scale and an empirical basis for examining how audience discourse varies across different video contexts within a major Indonesian digital science communication environment.

ACKNOWLEDGEMENTS

The author acknowledges the publicly accessible digital resources and open-source YouTube Data API v3 supported corpus construction, natural language processing, model development, and computational analysis especially the YouTube channel Kok Bisa?

REFERENCES

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7226–7249. https://doi.org/10.18653/v1/2022.acl-long.500

Austin, J. L. (1962). *How to do things with words*. Oxford University Press.

Barik, A. M., Mahendra, R., & Adriani, M. (2019). Normalization of Indonesian-English code-mixed Twitter data. *Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019)*, 417–424. https://doi.org/10.18653/v1/D19-5554

Bucchi, M., & Trench, B. (Eds.). (2021). *Routledge handbook of public communication of science and technology* (3rd ed.). Routledge.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dubovi, I., & Tabak, I. (2020). An empirical analysis of knowledge co-construction in YouTube comments. *Computers & Education, 156*, Article 103939. https://doi.org/10.1016/j.compedu.2020.103939

Dubovi, I., & Tabak, I. (2021). Interactions between emotional and cognitive engagement with science on YouTube. *Public Understanding of Science, 30*(6), 759–776. https://doi.org/10.1177/0963662521990848

Fauzan, M. F. D. (2026). *Kok Bisa? physics discourse analytics: An interactive research companion for the Indonesian Public Discourse Corpus (IPDC)* [Interactive web application]. GitHub Pages. Kok Bisa? Physics Discourse Analytics

Hill, V. M., Grant, W. J., McMahon, M. L., & Singhal, I. S. (2022). How prominent science communicators on YouTube understand the impact of their work. *Frontiers in Communication, 7*, Article 1014477. https://doi.org/10.3389/fcomm.2022.1014477

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Vosoughi, S., & Roy, D. (2016). Tweet acts: A speech act classifier for Twitter. *Proceedings of the International AAAI Conference on Web and Social Media, 10*(1), 711–714. https://doi.org/10.1609/icwsm.v10i1.14821