# Evaluating Authenticity in Digital Apology Discourse on Bluesky

Ryan Phua

rphua@g.ucla.edu

## Â§ 0 ABSTRACT

This study examines celebrity apology discourse on Bluesky, a decentralized social media platform, to identify factors correlating with positive versus negative sentiment in public accountability discussions. Using computational methods including VADER sentiment analysis, BERTopic topic modeling, and network analysis on 2,081 Bluesky posts, this research reveals that topic choice is the primary predictor of sentiment (F=12.45, p<0.001, Î·Â²=0.18), with authenticity-focused discourse receiving 68% positive sentiment compared to 28% for skepticism-focused discussions. Network analysis demonstrates that Bluesky's decentralized architecture creates fragmented discourse patterns (17 disconnected components) that differ significantly from centralized platforms like Twitter. Statistical validation confirms that specific language markers predict sentiment with large effect sizes (Cohen's d=0.52), suggesting that audiences distinguish between genuine accountability and performative apology. These findings contribute to understanding how platform architecture shapes accountability discourse and provide empirical evidence for factors that correlate with effective public apologies in digital spaces.

## Â§ 1 INTRODUCTION

Public apologies have become a defining feature of contemporary digital culture. When celebrities, influencers, or public figures face controversy, social media platforms serve as both the stage for their apologies and the arena where public judgment unfolds. The speed and visibility of these processes have transformed accountability from a private matter into a collective performance, where millions of users can instantly evaluate, critique, and amplify their responses. Understanding what makes an apology effective in this environment has implications for crisis communication, platform design, and our broader understanding of how digital spaces shape social norms around accountability.

This research examines apology discourse on Bluesky, a decentralized social media platform that emerged as an alternative to Twitter. By analyzing 2,081 posts discussing celebrity apologies, I investigate which factors correlate with positive versus negative sentiment in public accountability discussions. The central research question guiding this study is: What factors in Bluesky apology discourse correlate with positive versus negative sentiment, and how do these patterns inform our understanding of effective public accountability?

This question emerged from an initial interest in analyzing YouTube apology videos and their comment sections, but preliminary research revealed the value of first understanding discourse patterns on text-based platforms. Bluesky provides a unique research context because its decentralized architecture and smaller, more curated communities may create different dynamics than centralized platforms like Twitter or Facebook. The platform's emphasis on user control over content moderation and its growing user base of journalists, academics, and engaged citizens make it an ideal site for studying how informed audiences evaluate public apologies.

The significance of this research extends beyond understanding individual apologies. As cancel culture and call-out culture have become dominant modes of online accountability, questions about what constitutes genuine versus performative apology have taken on broader cultural importance. Understanding which apology strategies resonate with audiences can help distinguish between authentic accountability and strategic image management, with implications for how we think about justice, forgiveness, and power in digital spaces.

## Â§ 2 LITERATURE REVIEW

The question of what makes an apology effective has preoccupied communication scholars for decades, but the social media era has intensified this challenge by making audience responses immediately visible and quantifiable. Sandlin and Gracyalny's analysis of YouTube apologies demonstrates that perceived sincerity is the primary factor determining whether audiences forgive public figures, with 70% of comments focusing on the apologizer's reputation and character rather than the specific offense. Their finding that "reducing offensiveness" strategies correlate with perceptions of insincerity challenges conventional crisis communication advice and suggests that audiences have become sophisticated at detecting strategic image management.

This aligns with Kampf's earlier work identifying four categories of responsibility minimization tactics in public apologies: compromising the performative verb (saying "sorry" instead of "apologize"), blurring the offense's nature, questioning whether anyone was actually harmed, and questioning the identity of the offender. Together, these studies establish that audiences distinguish between authentic accountability and performative apology, and that this distinction drives their willingness to forgive.

Analyzing apology discourse at scale requires computational methods that can process large volumes of social media text while preserving meaningful patterns. Doogan and Buntine's systematic review of 189 studies finds that researchers overwhelmingly use Latent Dirichlet Allocation (LDA) despite its known limitations with short texts, and that there is a troubling disconnect between topic model development and the needs of applied researchers. Their critique that automated evaluation metrics like perplexity don't align with human interpretability directly informs my decision to use BERTopic, which leverages pre-trained language models to capture semantic meaning that traditional bag-of-words approaches miss.

The network analysis literature (Borgatti et al.; Newman; Wasserman and Faust) provides complementary methods for understanding how apology discourse spreads through social networks, revealing whether discussions form cohesive communities or fragmented clusters. My network analysis found that Bluesky apology discourse is highly fragmented with 17 disconnected components, suggesting that apologies function as isolated events rather than unified community conversations.

## Â§ 3 PLATFORM, TOPIC, AND DATA SELECTION

Bluesky was selected as the research platform for several strategic reasons. First, as a decentralized social media platform, Bluesky offers a unique architectural context that differs fundamentally from centralized platforms like Twitter or Facebook. This decentralization means that content moderation, algorithmic curation, and community norms operate differently, potentially creating distinct patterns in how accountability discourse unfolds.

Second, Bluesky's user base consists largely of early adopters, journalists, academics, and users seeking alternatives to Twitter's increasingly toxic environment. This demographic composition suggests a more engaged and potentially more critical audience for evaluating public apologies. Understanding how this specific community discusses accountability provides insights into how informed, media-literate audiences evaluate public figures' responses to controversy.

The dataset consists of 2,081 Bluesky posts collected in November 2024 using the Bluesky API. Posts were identified using keywords related to apology discourse, including "apology," "apologize," "sorry," "accountability," and related terms. The collection process prioritized posts that explicitly discussed celebrity or influencer apologies, filtering out casual uses of apology language in interpersonal contexts. All data was collected from public Bluesky posts in accordance with the platform's terms of service and API usage guidelines.

## Â§ 4 METHODS

This study employs an integrated computational approach combining four complementary methods: sentiment analysis (VADER), topic modeling (BERTopic), network analysis (graph theory), and statistical validation. Each method provides unique insights that, when integrated, enable comprehensive understanding of factors correlating with sentiment in apology discourse.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool specifically designed for social media text. Each post was analyzed using VADER to generate four scores: positive proportion, negative proportion, neutral proportion, and compound score (normalized from -1 to +1). Posts were classified as positive (compound â‰¥ 0.05), negative (compound â‰¤ -0.05), or neutral (-0.05 < compound < 0.05) based on established VADER thresholds.

BERTopic is a neural topic modeling approach that uses transformer-based language models (BERT) to generate contextual embeddings, then applies clustering algorithms to identify topics. Unlike traditional LDA, BERTopic captures semantic meaning and handles short texts effectively. The BERTopic pipeline consisted of text preprocessing, embedding generation using pre-trained BERT models, dimensionality reduction with UMAP, clustering with HDBSCAN, and topic representation extraction using c-TF-IDF.

Network analysis examines relationships between entities (nodes) through their connections (edges). This study constructed two networks: a mention network (directed graph where nodes are users and edges represent mentions) and a hashtag network (bipartite graph connecting users to hashtags they use). For each network, I calculated degree centrality, betweenness centrality, connected components, network density, and clustering coefficient.

Statistical validation ensures that findings are robust and not artifacts of sampling or measurement error. This study employed chi-square tests, ANOVA, t-tests, effect size calculations (Cohen's d, Î·Â²), and confidence intervals. All statistical tests used Î± = 0.05 significance level.

## Â§ 5 FINDINGS

The sentiment analysis revealed a more nuanced landscape than popular narratives about "cancel culture" would suggest. Contrary to assumptions about overwhelmingly negative online discourse, 51% of posts expressed positive sentiment, 43% negative, and 6% neutral. Statistical validation confirmed this distribution is significant (Ï‡Â²=45.3, p<0.001), meaning it differs substantially from what would be expected by chance. This finding suggests that Bluesky users are more receptive to apologies than stereotypes about online discourse would predict.

The integration of topic modeling with sentiment analysis revealed that topic choice explains 18% of sentiment variance (F=12.45, p<0.001, Î·Â²=0.18), with authenticity-focused topics receiving 68% positive sentiment compared to 28% for skepticism-focused topics. ANOVA testing confirmed that topic significantly predicts sentiment, with a large effect size indicating practical as well as statistical significance. This finding validates theoretical work showing that perceived sincerity drives audience responses. Bluesky users distinguish between authentic accountability and performative apology, and this distinction strongly correlates with sentiment.

Network analysis revealed striking differences from centralized platforms like Twitter. The mention network showed 17 connected components, indicating 17 separate conversation clusters with no connections between them. Network density was low (0.08), meaning only 8% of possible connections exist. Hashtag usage was minimal (12.5%), far lower than typical Twitter discourse. This fragmentation pattern suggests that Bluesky's decentralized architecture creates fundamentally different discourse dynamics than centralized platforms, potentially reducing mob dynamics while enabling diverse evaluations.

Analysis of word usage patterns revealed that certain terms consistently appear in positive versus negative sentiment posts. Positive sentiment markers include "genuine," "sincere," "accountability," "growth," "learning," and "change." Negative sentiment markers include "performative," "PR stunt," "damage control," "fake," "insincere," and "calculated." The large effect size (Cohen's d=0.52) indicates that these language markers are strong predictors of sentiment, not merely weak associations.

## Â§ 6 DISCUSSION

This research demonstrates that digital accountability discourse is more nuanced than popular narratives suggest. While "cancel culture" rhetoric often portrays online audiences as uniformly hostile, the data reveals that Bluesky users are receptive to genuine apologies while skeptical of performative ones. This distinction matters because it suggests that effective accountability is possible in digital spacesâ€”public figures who demonstrate authentic accountability can shift sentiment positively.

The finding that topic choice is the primary predictor of sentiment has important implications for how we understand digital discourse. Rather than treating all apology discussions as equivalent, we must recognize that what people discuss matters more than how much they discuss. Authenticity-focused discourse creates different dynamics than skepticism-focused discourse, and these differences correlate with measurably different sentiment outcomes.

The network fragmentation finding challenges assumptions from Twitter research and demonstrates that platform architecture fundamentally shapes discourse patterns. Bluesky's decentralization creates isolated conversations that may reduce mob dynamics while enabling diverse evaluations. This has implications for platform design: if we want to create spaces where nuanced accountability discussions can occur, decentralized architectures may be more effective than centralized ones.

For public figures and communications professionals, this research provides evidence-based guidance for crafting effective apologies: emphasize authenticity using language that signals genuine accountability rather than strategic image management, be specific with concrete actions and acknowledgments, avoid minimization tactics that reduce responsibility, and consider platform-specific dynamics when tailoring responses.

## Â§ 7 LIMITATIONS AND FUTURE RESEARCH

Several limitations should be considered when interpreting these findings. The dataset represents a snapshot from November 2024, and discourse patterns may vary over time as platform norms evolve and different controversies emerge. Findings are specific to Bluesky's user base and architecture, and patterns may differ on other platforms with different demographics and structural features. The dataset includes only public posts from users who chose to discuss apologies, meaning silent audiences and private conversations are not captured. This research identifies correlations, not causal relationshipsâ€”we cannot conclude that using authenticity language causes positive sentiment, only that they correlate.

Future research directions include cross-platform comparison analyzing the same apologies across multiple platforms (YouTube, Twitter, Bluesky) to reveal how platform architecture shapes discourse, temporal analysis tracking sentiment evolution over longer time periods, video analysis extending this research to analyze apology videos themselves, experimental validation testing whether manipulating authenticity markers causally influences sentiment, and demographic analysis examining how different user demographics evaluate apologies.

## Â§ 8 CONCLUSION

This research demonstrates that digital accountability discourse is more nuanced and analyzable than popular narratives suggest. By integrating sentiment analysis, topic modeling, network analysis, and statistical validation, the study reveals that topic choice is the primary factor correlating with sentiment in Bluesky apology discourse, with authenticity-focused discussions receiving substantially more positive sentiment than skepticism-focused ones. The finding that Bluesky's decentralized architecture creates fragmented discourse patterns challenges assumptions from Twitter research and demonstrates that platform design fundamentally shapes how accountability conversations unfold.

Most fundamentally, this research challenges the narrative that online audiences are uniformly hostile toward public figures who apologize. The data reveals that Bluesky users are receptive to genuine accountability while skeptical of performative apologyâ€”a distinction that suggests effective accountability is possible in digital spaces when public figures demonstrate authentic commitment to growth and change. As digital platforms continue to shape how we negotiate questions of justice, forgiveness, and power, understanding these dynamics becomes increasingly important for creating online spaces that enable meaningful accountability rather than merely performative gestures.

## Â§ 9 REFERENCES

Benoit, William L. Accounts, Excuses, and Apologies: A Theory of Image Restoration Strategies. State University of New York Press, 1995.

Borgatti, Stephen P., et al. Analyzing Social Networks. 2nd ed., SAGE Publications, 2018.

Clark, Meredith D. &quot;DRAG THEM: A Brief Etymology of So-called 'Cancel Culture.'&quot; Data & Society Research Institute, 2020.

Doogan, Caitlin, and Wray Buntine. &quot;A Systematic Review of the Use of Topic Models for Short Text

Social Media Analysis.&quot; Artificial Intelligence Review, vol. 56, 2023, pp. 14223-14255.

Kampf, Zohar. &quot;Public (Non-) Apologies: The Discourse of Minimizing Responsibility.&quot; Journal of Pragmatics, vol. 41, no. 11, 2009, pp. 2257-2270.

Newman, Mark E. J. Networks. 2nd ed., Oxford University Press, 2018.

Nugroho, Rahmad, et al. &quot;A Survey of Recent Methods on Deriving Topics from Twitter: Algorithm to Evaluation.&quot; Knowledge and Information Systems, vol. 62, 2020, pp. 2485-2519.

Sandlin, Jean Kelso, and Monica L. Gracyalny. &quot;Seeking Sincerity, Finding Forgiveness: YouTube Apologies as Image Repair.&quot; Public Relations Review, vol. 44, no. 3, 2018, pp. 393-406.

Wasserman, Stanley, and Katherine Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

Zhao, He, et al. &quot;Topic Modelling Meets Deep Neural Networks: A Survey.&quot; Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-21), 2021.