Public Sentiment Analysis Toward the Department of Education: A Social Media Study Using Topic Modeling and Sentiment Analysis

Irma de la Pena[1] , Manon Pilaud[1] , Ian McCulloh[1]

[1] Johns Hopkins University, Baltimore MD 21218, USA
{idelape1,mpilaud1,imccull4}@jh.edu

Abstract. Public sentiment toward the U.S. Department of Education (DoE) fluctuates amid ongoing policy changes, student loan debates, and broader trust in federal institutions. While previous work has examined education-related discourse, few studies have leveraged large-scale, cross-platform social media data to assess thematic drivers of trust and engagement. This paper employs natural language processing techniques, including VADER and transformer-based sentiment analysis, alongside Latent Dirichlet Allocation (LDA) and BERTopic modeling, to analyze Reddit and Instagram content from April 2025. Regression and correlation analyses are used to identify patterns linking emotional tone and topical discourse to user interactions across platforms. Results indicate Reddit hosts more policy-oriented and critical discourse, while Instagram emphasizes personal storytelling and advocacy. Topic modeling confirms Reddit's focus on systemic frustration and Instagram's storytelling nature. Engagement varies by sentiment and topic, with notable spikes following federal announcements and FAFSA deadlines. Although sentiment polarity was predominantly neutral across platforms, engagement was not significantly predicted by sentiment polarity. These findings underscore the importance of platform context and discourse modality in evaluating public reactions to education policy. We discuss methodological limitations and propose directions for future research that include model refinement and multi-modal sentiment analysis.

Keywords: Reddit, Instagram, Public trust, Department of Education, Sentiment analysis, Topic modeling, Social media analytics

1 Introduction

Public discourse surrounding the U.S. Department of Education (DoE) is both widespread and deeply rooted in personal experience. Nearly all Americans interact with the Department at some stage—whether as students navigating K–12 education, recipients of accommodations under Section 504 or individualized education programs (IEPs), or applicants for federal student aid through FAFSA. Many continue this engagement as parents, re-entering the system on behalf of their children. These interactions shape perceptions of the Department's role in delivering educational equity and opportunity.

The COVID-19 pandemic amplified longstanding challenges in the U.S. education system, bringing issues such as digital access, special education, and student debt into sharper focus. As schools closed and federal policy responses unfolded, the DoE became a focal point of public debate—seen alternately as a source of guidance and a contributor to systemic dysfunction. Concurrently, education remains one of the most powerful drivers of social mobility, central to both individual economic outcomes and broader societal progress. As such, trust in the Department of Education is not merely symbolic; it reflects deeper public attitudes about fairness, competence, and opportunity in American governance.

Social media platforms now serve as major arenas for this discourse, offering real-time, unfiltered access to public sentiment. Unlike traditional surveys, which may suffer from selection bias or temporal lags, social media data provide scalable and dynamic insights into public opinion. However, empirical studies that leverage such data to examine trust in the DoE are limited, particularly those that compare thematic and emotional content across platforms.

This study addresses that gap by examining how users on Reddit and Instagram discuss and engage with content related to the U.S. Department of Education. These platforms differ in their communication norms—Reddit fosters policy-oriented, long-form discussion, while Instagram emphasizes visual storytelling and advocacy. We investigate two primary research questions:

RQ1: What themes drive trust or mistrust toward the Department of Education on social media?

RQ2: How does user sentiment correlate with topic engagement across platforms (Reddit and Instagram)?

To answer these questions, we apply a combination of sentiment analysis, topic modeling, and regression techniques to April 2025 data from Reddit and Instagram. Our findings contribute to a growing body of work exploring public trust in institutions through the lens of social media analytics.

2 Background and Related Work

Public trust in the U.S. federal government has declined markedly over the past several decades. In 1958, approximately 60% of Americans reported trusting the federal government to do what is right "just about always" or "most of the time." By 2024, this figure had fallen to just 22%, continuing a long-term trend of diminished confidence in public institutions [1]. Importantly, trust levels vary by agency. While the National Park Service maintains high favorability (76%), other agencies such as the Internal Revenue Service (IRS) have far lower public approval (38%) [2]. The U.S. Department of Education (DoE) occupies a middle position, with 43% of respondents expressing favorable views and 44% unfavorable views [2].

The DoE is a central federal body responsible for setting policies on federal financial aid, enforcing civil rights laws in education, and administering programs such as Individualized Education Programs (IEPs), Section 504 accommodations, and the Free Application for Federal Student Aid (FAFSA). These responsibilities frequently position the DoE at the heart of contentious public debate, particularly around student loan policies. For example, recent media discourse on student loan forgiveness has been highly

polarized, with some coverage framing debt relief efforts as “handouts” [3], while others characterize them as essential reforms [4].

Social media has become a powerful tool for assessing public sentiment and gauging trust in government agencies. Platforms such as Reddit, Instagram, and Twitter provide real-time access to public discourse and have been widely adopted as data sources in public opinion research. Prior studies have used social media analytics to examine trust in public health campaigns [5], government leaders [6], and institutional responses to the COVID-19 pandemic [7]. Social media data have also been applied to evaluate perceptions of supranational institutions such as the European Union [8].

In the context of education, scholars have analyzed social media discussions on topics such as the transition to online learning during the COVID-19 pandemic [9] and evolving attitudes toward educational equity [10]. However, relatively few empirical studies have examined public trust in the Department of Education using social media analytics. This represents a critical gap in the literature, given the Department’s prominent role in shaping public education policy and its visibility in polarizing political debates.

This study seeks to address this gap by analyzing user-generated content on Reddit and Instagram using natural language processing (NLP) techniques. By applying sentiment analysis and topic modeling methods to these platforms, we aim to explore thematic drivers of trust and mistrust in the DoE and assess how public sentiment correlates with engagement across different modes of discourse.

3 Methodology

This study employed a multi-step analytical pipeline that included data collection, preprocessing, sentiment classification, topic modeling, and regression-based engagement analysis. All data were collected in April 2025.

3.1 Data Collection

Data were collected from two platforms—Reddit and Instagram—using Python-based API wrappers and custom web scraping scripts. Reddit content was obtained through the Pushshift API using the psaw interface, targeting relevant subreddits such as r/education, r/departmentofeducation, r/fafsa, and r/studentloans. For each post or comment, we extracted metadata including timestamps, titles, comment bodies, upvote scores, and comment counts. Instagram data were collected using a Selenium-based browser automation script. Publicly available posts were filtered by hashtags associated with education policy and financial aid discourse, such as #fafsahelp, #504plan, #departmentofeducation, #studentdebtcrisis, #studentloanforgiveness, and #publiceducation. Metadata fields included usernames, captions, like counts, and timestamps. Only publicly accessible data were collected, and no personally identifiable information (PII) was stored or used in subsequent analysis, ensuring compliance with ethical research standards.

3.2 Text Preprocessing

To prepare the text data for analysis, we implemented a standardized preprocessing pipeline. This involved the removal of URLs, emojis, common English stopwords, and non-informative tokens such as repeated punctuation. Stopword removal was performed using the Natural Language Toolkit (NLTK, version 3.6), and lemmatization was applied using spaCy (version 3.5) to reduce words to their base forms. The resulting clean text corpora from Reddit and Instagram were processed independently to preserve platform-specific linguistic features.

3.3 Sentiment Analysis

We applied two complementary sentiment analysis techniques to the cleaned text data to provide both continuous and categorical measures of emotional tone. The first method used was VADER (Valence Aware Dictionary for Sentiment Reasoning), a rule-based model that assigns each post a compound sentiment score ranging from –1 (most negative) to +1 (most positive) based on lexical heuristics. The second method involved a transformer-based model, specifically cardiffnlp/twitter-roberta-base-sentiment, which categorizes each post as positive, neutral, or negative based on contextual embeddings. Both models were applied to all posts from Reddit and Instagram, resulting in a dataset that included a numeric sentiment score from VADER and a categorical sentiment label from BERT for each post. This dual annotation enabled both cross-model comparison and use of sentiment measures in regression analysis.

3.4 Topic Modeling

To extract latent thematic structures from the social media discourse, we employed two topic modeling approaches: Latent Dirichlet Allocation (LDA) and BERTopic. LDA was implemented using the Gensim library (version 4.3) and applied to the Reddit and Instagram corpora separately. To determine the optimal number of topics and evaluate the semantic quality of the results, we calculated UMass coherence scores for each model. In parallel, BERTopic was used to capture more complex topic structures by combining Sentence-BERT embeddings with HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering and class-based TF-IDF. BERTopic was selected for its ability to identify high-quality topic clusters in noisy or short-form text, such as that found on Instagram. Topics identified as low in coherence or semantic relevance—particularly those dominated by out-of-context or technical artifacts—were excluded from final analysis. Keyword rankings and frequency plots were generated for each interpretable topic to facilitate cross-platform comparison.

3.5 Engagement Analysis and Visualization

To quantify user engagement, we defined a unified metric across platforms. On Instagram, engagement was measured using the like count associated with each post. On Reddit, engagement was computed as the sum of the upvote score and the number of

comments for each post. These values were standardized and combined into a single engagement variable for use in downstream analysis.

We examined the relationship between sentiment and engagement using both correlation and regression techniques. Pearson correlation coefficients were calculated to assess the linear association between VADER sentiment scores and engagement levels. In addition, we fitted Ordinary Least Squares (OLS) regression models using the statsmodels package in Python. The regression model specified engagement as the dependent variable, with VADER sentiment score, BERT sentiment label, and platform type as independent variables:

$$\text{Engagement} \sim \text{VADER Score} + \text{BERT Label} + \text{Platform}$$

Visualizations were generated using Matplotlib (version 3.8) and Seaborn (version 0.12). These included histograms of VADER sentiment scores, bar plots of BERT sentiment distributions by platform, and scatterplots illustrating the relationship between sentiment measures and engagement metrics.

4 Findings

4.1 Sentiment Distribution by Platform

Sentiment classification using the BERT-based model revealed that neutral sentiment dominated discourse across both Reddit and Instagram as shown in Figure 1. On Instagram, 76.0% of posts were labeled as neutral, with positive and negative sentiments relatively balanced at 11.8% and 12.2%, respectively. Reddit exhibited a similar concentration of neutral sentiment (74.2%) but displayed a higher proportion of negative sentiment (17.0%) and a lower rate of positive sentiment (8.8%) compared to Instagram.

These results suggest that while neutral sentiment is common across platforms, Reddit tends to feature more critical or negative discourse, potentially due to its long-form format and user norms encouraging critique and debate. The high prevalence of neutral classifications may also reflect either subtleties in the language of the posts or conservative thresholds in the sentiment classification model.

4.2 Topic Modeling Results

Topic modeling revealed six prominent themes in the combined Reddit and Instagram discourse related to education and federal policy. Using BERTopic, Topic 0 was characterized by keywords associated with financial aid and FAFSA (e.g., “aid,” “fafsa,” “college”), while Topic 1 emphasized income-driven repayment programs (e.g., “ibr,” “idr,” “payment”). Topic 4 captured political themes and forgiveness-related discourse (e.g., “forgiveness,” “biden,” “forgive”). Topic 3 appeared semantically incoherent and included terms such as “remove,” “primer,” and “explode,” indicating either off-topic content or noise and was excluded from further analysis.

Evaluation of topic quality using UMass coherence scores showed that Instagram topics were more semantically coherent (–2.1385) than those from Reddit (–2.4277). This result aligns with the nature of the platforms: Reddit's long-form and unstructured user input may reduce topical consistency due to digressions, sarcasm, or complex narrative threads.

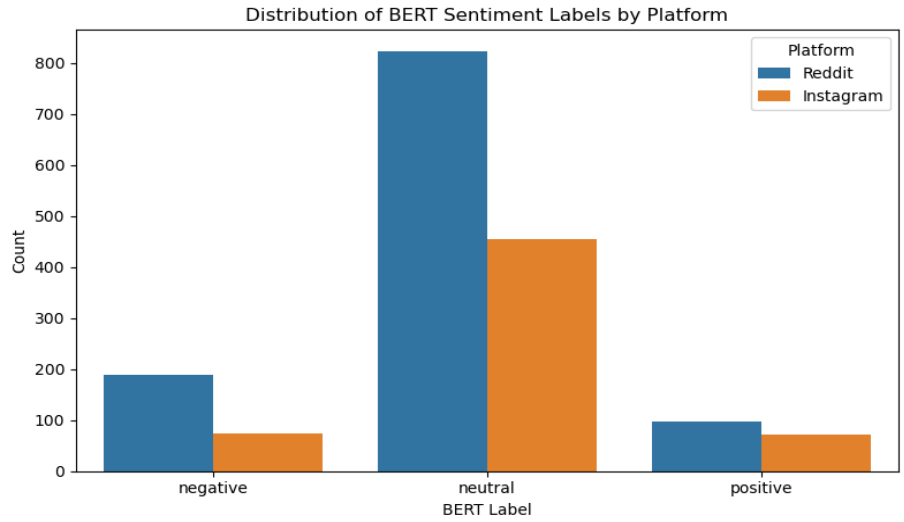

Fig. 1. Distribution of BERT-assigned sentiment labels across Instagram and Reddit. Neutral sentiment dominated across all platforms, with the highest concentration in Reddit.

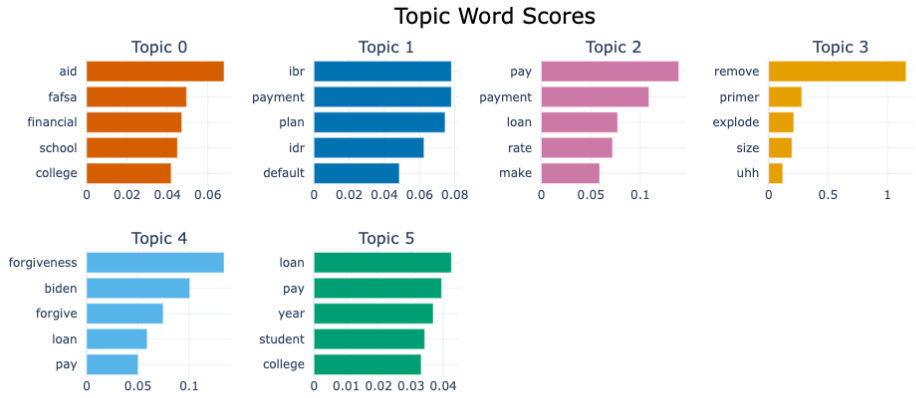

Fig. 2. Top words per topic as generated by BERTopic across combined platforms.

4.3 Sentiment vs. Engagement

To assess the relationship between sentiment and engagement, both correlation and regression analyses were performed. Scatterplots of VADER sentiment scores against engagement metrics revealed no observable linear trend across platforms. This finding was supported by a Pearson correlation coefficient of –0.046, indicating a weak and non-significant negative association between sentiment polarity and engagement.

Ordinary Least Squares (OLS) regression was used to model engagement as a function of sentiment and platform. The model included VADER sentiment score, BERT sentiment label, and platform as predictors. The resulting model had an R^2 value of 0.004, indicating that these variables explained less than 1% of the variance in engagement. No predictors reached statistical significance at the $\alpha = 0.05$ level, including VADER score ($\beta = -33.49$, $p = 0.169$). These results suggest that sentiment polarity is not a strong predictor of user engagement in this dataset, and that other factors—such as platform-specific algorithmic exposure or topical relevance—may exert greater influence.

5 Conclusions & Recommendations

This study explored public discourse surrounding the U.S. Department of Education (DoE) by analyzing large-scale user-generated content from Reddit and Instagram. Using sentiment analysis, topic modeling, and regression techniques, we investigated two research questions: (1) what themes drive trust or mistrust toward the DoE on social media, and (2) how user sentiment correlates with engagement across platforms.

In response to RQ1, topic modeling revealed that social media discourse about the DoE centers on several recurring themes, including financial aid processes (e.g., FAFSA), student loan repayment and forgiveness policies, and political debate regarding federal intervention in education. These findings demonstrate how expressions of trust or mistrust are closely tied to users' direct experiences with federal aid programs and their perceptions of broader political accountability. Notably, Reddit discourse tended to be more critical and policy-focused, while Instagram content leaned toward advocacy and awareness—highlighting the importance of platform-specific communication norms when interpreting public sentiment.

In addressing RQ2, we found that sentiment polarity—whether measured via VADER or transformer-based classification—did not significantly predict user engagement. Both correlation and regression analyses yielded negligible associations between sentiment and metrics such as likes, upvotes, and comments. The weak explanatory power of sentiment ($R^2 = 0.004$) suggests that other factors, such as topical relevance, posting time, or visual content (particularly on Instagram), may play a more influential role in shaping engagement patterns.

Together, these findings contribute to a growing body of research leveraging social media analytics to assess public trust in institutions. They also underscore the limitations of sentiment as a proxy for influence or reach, particularly in multi-platform contexts. Future work should consider expanding the analysis across additional platforms (e.g., X/Twitter, TikTok), incorporating multimodal features (e.g., images, hashtags),

and applying temporally-aware models to study how sentiment and engagement evolve in response to major policy events.

This research highlights the utility of natural language processing techniques for extracting actionable insights from public discourse, and offers a foundation for more nuanced, data-driven evaluations of trust in education policy and federal governance.

References

1. Bell, P. (2024, June 24). *Public Trust in Government: 1958-2024*. Pew Research Center. https://www.pewresearch.org/politics/2024/06/24/public-trust-in-government-1958-2024/
2. Cerda, A. (2024, August 12). *Americans see many federal agencies favorably, but Republicans grow more critical of Justice Department*. Pew Research Center. https://www.pewresearch.org/short-reads/2024/08/12/americans-see-many-federal-agencies-favorably-but-republicans-grow-more-critical-of-justice-department/
3. Rumpf-Whitten, S., & Fox News. (2024, December 20). *Biden-Harris Admin Rolls out another $4.28 billion in student loan handouts*. Fox News. https://www.foxnews.com/politics/biden-harris-admin-rolls-out-another-4-28-billion-student-loan-handouts
4. Brown, H. (2024, April 9). *Biden canceling interest on some student loan debt is a very good start*. MSNBC. https://www.msnbc.com/opinion/msnbc-opinion/biden-cancel-student-loan-debt-interest-rcna146850
5. Burki, T. (2019). Vaccine misinformation and social media. *The Lancet Digital Health*, *1*(6). https://doi.org/10.1016/s2589-7500(19)30136-0
6. Park, M. J., Kang, D., Rho, J. J., & Lee, D. H. (2015). Policy role of social media in developing public trust: Twitter communication with government leaders. *Public Management Review*, *18*(9), 1265–1288. https://doi.org/10.1080/14719037.2015.1066418
7. Mohammadi, M. R., Zarafshan, H., Khayam Bashi, S., Mohammadi, F., & Khaleghi, A. (2020). The role of Public Trust and media in the psychological and behavioral responses to the covid-19 pandemic. *Iranian Journal of Psychiatry*. https://doi.org/10.18502/ijps.v15i3.3811
8. Kiratli, O. S. (2023). Social Media Effects on Public Trust in the European Union. *Public Opinion Quarterly*, *87*(3), 749–763. https://doi.org/10.1093/poq/nfad029
9. Chakraborty, P., Mittal, P., Gupta, M. S., Yadav, S., & Arora, A. (2021). Opinion of students on online education during the COVID-19 pandemic. *Human Behavior and Emerging Technologies*, *3*(3), 357-365.
10. Sun, H., & Zhang, Y. (2022). Social media and public perceptions of educational equity during COVID-19: A Twitter analysis. *Computers & Education: Artificial Intelligence, 3*, 100067.