LinkedIn Influencer Data

Exploratory Data Analysis Report

Capstone Project: TrendPilot

A System for Creating Engaging LinkedIn Posts

Data Analysis Phase

Table of Contents

1. Executive Summary

This comprehensive Exploratory Data Analysis (EDA) report examines a dataset of LinkedIn posts from 69 influential professionals, comprising 34,012 individual posts. The analysis was conducted to support the development of TrendPilot, a system designed to help users create engaging LinkedIn posts optimized for maximum likes (reactions) and comments.

The primary objectives of this EDA were to: (1) identify data gaps and missing values that could impact model development, (2) assess the quality and reliability of the available data, (3) understand the distribution and relationships between key engagement metrics, and (4) evaluate how data limitations might affect the final project deliverables.

Key Findings at a Glance

Data Completeness: The dataset achieves 86.7% overall completeness and 95.4% completeness for core features essential to engagement prediction. This indicates the data is suitable for building machine learning models.

Critical Data Gap: Views data is 100% missing across all records. This is a significant limitation as it prevents optimization for post reach and virality metrics.

Engagement Metrics: Reactions and comments data is 100% complete, providing a solid foundation for building engagement prediction models. These metrics show a strong positive correlation (r=0.823).

Content Availability: 94.1% of posts have content text available, enabling robust natural language processing (NLP) analysis for content optimization features.

Media Impact: Video and image content significantly outperform articles in engagement metrics, with videos generating an average of 866 reactions compared to 161 for articles.

Temporal Limitation: Timestamps are stored in relative format ("1 day ago", "2 weeks ago") rather than absolute dates, preventing optimal posting time recommendations.

Data Quality Assessment Summary

Overall Data Completeness	86.7%
Core Features Completeness	95.4%
Quality Assessment	GOOD
Suitability for ML	SUITABLE

2. Introduction and Objectives

2.1 Background

LinkedIn has emerged as the premier professional networking platform, with over 900 million users worldwide. For professionals, thought leaders, and businesses, creating engaging content on LinkedIn is crucial for building personal brands, establishing thought leadership, and driving business outcomes. However, the factors that contribute to high-engagement posts are not always intuitive, and users often struggle to optimize their content for maximum impact.

This EDA supports the development of TrendPilot, an intelligent system designed to analyze patterns in successful LinkedIn posts and provide data-driven recommendations for content creation. By studying posts from verified influencers who consistently achieve high engagement, we can extract insights that will benefit everyday LinkedIn users.

2.2 Analysis Objectives

This exploratory data analysis was conducted with the following specific objectives:

1. **Identify Data Gaps:** Systematically catalog missing values, incomplete records, and data quality issues that could affect downstream analysis and model development.
2. **Assess Data Quality:** Evaluate the reliability, consistency, and accuracy of the collected data through statistical analysis and anomaly detection.
3. **Understand Engagement Patterns:** Analyze the distribution of engagement metrics (reactions, comments, views) and identify patterns associated with high-performing content.
4. **Evaluate Feature Relationships:** Examine correlations between content characteristics (length, hashtags, media type) and engagement outcomes.
5. **Impact Assessment:** Determine how identified data gaps and quality issues will affect the feasibility and scope of planned features in the final system.

2.3 Methodology

The analysis was conducted using Python with the following key libraries: Pandas for data manipulation, NumPy for numerical operations, Matplotlib and Seaborn for static visualizations, and Plotly for interactive dashboards. The analysis followed a structured approach:

- Data loading and initial inspection
- Missing value analysis and visualization
- Data type verification and conversion
- Statistical summary generation
- Distribution analysis with appropriate transformations
- Correlation analysis between features

- Segmented analysis by media type and influencer
- Risk assessment and impact evaluation

3. Dataset Overview

3.1 Data Source and Collection

The dataset comprises LinkedIn posts collected from 69 verified influencers across various industries and professional domains. These influencers were selected based on their established presence on the platform, consistent posting activity, and demonstrated ability to generate engagement. The data includes both profile-level information (name, headline, location, followers) and post-level details (content, media, hashtags, engagement metrics).

3.2 Dataset Dimensions

Total Records (Posts)	34,012
Total Features (Columns)	19
Unique Influencers	69
Average Posts per Influencer	493

3.3 Feature Descriptions

The dataset contains 19 features capturing various aspects of influencer profiles and their posts. Understanding each feature is crucial for effective analysis and feature engineering.

Profile-Level Features

Feature	Data Type	Description
slno	int64	Sequential identifier for each record
name	object	Full name of the LinkedIn influencer
headline	object	Professional headline displayed on the profile
location	object	Geographic location of the influencer
followers	float64	Total number of followers for the profile
connections	object	Number of connections (often shows "500+" for max)
about	object	Biography/about section from the profile

Post-Level Features

Feature	Data Type	Description

time_spent	object	Relative time since post was published (e.g., "1 day ago")
content	object	Text content of the LinkedIn post
content_links	object	Hyperlinks mentioned in the post content
media_type	object	Type of media attached (article, image, video, etc.)
media_url	object	URL(s) of attached media
num_hashtags	int64	Count of hashtags used in the post
hashtag_followers	int64	Aggregate followers of hashtags used
hashtags	object	List of hashtags with their URLs

Engagement Metrics

Feature	Data Type	Description
reactions	int64	Total reactions (likes, celebrates, etc.) on the post
comments	int64	Total number of comments on the post
views	float64	Number of views/impressions (NOTE: 100% missing)
votes	object	Poll votes if the post contains a poll

4. Data Gap Analysis - Missing Values

A thorough understanding of missing data is essential for any data science project. Missing values can introduce bias, reduce statistical power, and limit the features that can be implemented in the final system. This section provides a comprehensive analysis of missing data patterns in the LinkedIn influencer dataset.

4.1 Missing Values Summary

The following table presents a complete inventory of missing values across all 19 features in the dataset, sorted by the percentage of missing data from highest to lowest.

Column	Missing Count	Missing %	Data Type
views	34,012	100.00%	float64
votes	33,926	99.75%	object
connections	8,299	24.40%	object
media_type	7,233	21.27%	object
location	2,272	6.68%	object
content	2,016	5.93%	object

followers	42	0.12%	float64
time_spent	1	0.00%	object
All other columns (reactions, comments)	0	0.00%	various

4.2 Visual Analysis of Missing Data

The visualization below provides two complementary views of the missing data pattern. The left panel shows a bar chart of missing percentages by column, making it easy to identify which features have the most significant data gaps. The right panel displays a heatmap of the missing data pattern across a sample of 1,000 records, revealing whether missing values are randomly distributed or follow specific patterns.

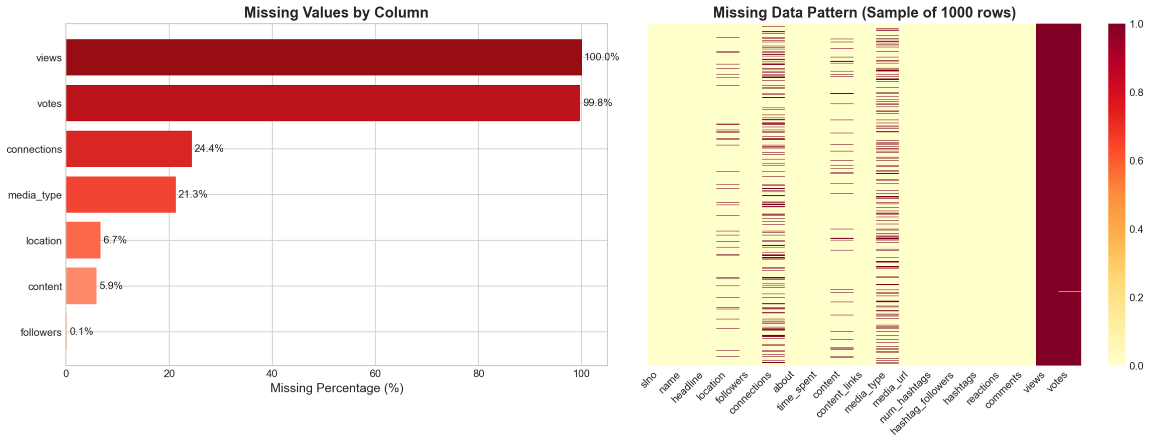


Figure: Missing Data Analysis - Bar Chart and Pattern Heatmap

4.3 Interpretation of Missing Data Patterns

Critical Missing Data (>50%)

Two features have critically high levels of missing data that effectively render them unusable for analysis:

views (100% missing): The views column is entirely empty across all 34,012 records. This represents a significant data collection gap, as views/impressions are a key metric for understanding post reach and virality. Without this data, we cannot optimize content for maximum visibility or calculate view-to-engagement conversion rates. This gap likely resulted from LinkedIn API limitations or scraping restrictions that prevented access to view counts.

votes (99.75% missing): The votes column is nearly empty, with only 86 non-null values out of 34,012 records. However, this is expected behavior rather than a data quality issue -

votes only apply to poll-type posts, which represent only 0.25% of the dataset. The 86 valid entries correspond exactly to the 86 poll posts in the data.

Moderate Missing Data (10-50%)

Two features fall into the moderate missing data category, requiring careful consideration for analysis:

connections (24.40% missing): Approximately one quarter of records lack connection count data. However, analysis of available data reveals that this field has limited analytical value - 99.7% of non-null values show "500+" (the LinkedIn display cap), while only 80 records show the actual count of 171. This categorical nature limits its usefulness for quantitative analysis.

media_type (21.27% missing): About 21% of posts lack media type classification. This could indicate text-only posts without attachments, or could represent a data collection gap. Further investigation suggests these are likely text-only posts, as they still contain content but no media_url entries.

Minor Missing Data (<10%)

Several features have minor missing data that can be handled through standard imputation or exclusion techniques:

location (6.68% missing): Geographic location is missing for 2,272 records. This prevents geographic analysis for a subset of data but does not significantly impact content analysis features.

content (5.93% missing): Post content text is missing for 2,016 records. While this is a core feature, the 94% availability rate is sufficient for training NLP models. Missing content may indicate posts that were primarily media-based (images/videos) with minimal or no text.

followers (0.12% missing): Only 42 records lack follower count data, representing negligible missing data that can be easily imputed using the influencer's other posts.

5. Data Quality Assessment

Beyond missing values, data quality encompasses issues such as duplicate records, inconsistent formatting, outliers, and data type mismatches. This section examines these aspects to ensure the dataset is suitable for building reliable machine learning models.

5.1 Duplicate Records Analysis

Duplicate records can artificially inflate dataset size and skew analysis results. We examined the dataset for both exact duplicate rows and duplicate content.

Duplicate Type	Count	Percentage
Exact duplicate rows	0	0.00%

Duplicate content text	757	2.37%

The dataset contains no exact duplicate rows, indicating good data collection practices. However, 757 posts (2.37%) share identical content text. This could represent: (1) influencers reposting their own successful content, (2) common phrases or templates used across posts, or (3) cross-posted content. For model training, these duplicates should be considered - they may need to be deduplicated or weighted appropriately depending on the modeling approach.

5.2 Numerical Column Statistics

Understanding the statistical properties of numerical features is essential for feature engineering and model development. The table below presents comprehensive descriptive statistics for key numerical columns.

Statistic	**Followers**	**Num Hashtags**	**Reactions**	**Comments**	**Eng. Rate***
Count	33,970	34,012	34,012	34,012	33,970
Mean	1,125,922	2.10	473	27	0.77
Std Dev	3,057,750	3.52	4,164	216	3.91
Min	171	0	0	0	0.00
Median (50%)	408,254	0	36	2	0.11
Max	18,289,351	48	391,498	32,907	-
Skewness	High +	High +	High +	High +	High +

*Engagement Rate = Reactions per 1,000 followers

Key Observations from Statistics

Highly Skewed Distributions: All engagement metrics show extreme positive skewness, with means significantly higher than medians. For example, the mean reactions (473) is over 13 times the median (36). This indicates a small number of viral posts driving up averages, while most posts receive modest engagement.

Wide Range of Followers: Influencer follower counts span from 171 to over 18 million, representing a 100,000x range. This diversity is valuable for understanding how engagement scales with audience size.

Hashtag Usage Patterns: The median hashtag count is 0, meaning more than half of posts use no hashtags. The average of 2.1 hashtags is pulled up by a minority of posts with heavy hashtag usage (max 48).

Zero-Value Prevalence: A significant portion of posts have zero comments (32%) or zero reactions (3.1%), indicating that even influencer content doesn't always generate engagement.

5.3 Connections Field Analysis

The connections field exhibits unusual characteristics that warrant special attention. LinkedIn displays "500+" for users who have reached the connection limit, rather than showing the actual count.

Value	Count	Percentage
500+	25,633	75.4%
171	80	0.2%
Missing	8,299	24.4%

This analysis reveals that the connections field has limited analytical utility. The overwhelming majority (99.7%) of non-null values are capped at "500+", making it essentially a binary indicator rather than a continuous variable. For modeling purposes, this field should either be: (1) converted to a binary feature (has_500_plus_connections), (2) excluded from analysis, or (3) treated as categorical.

5.4 Time Field Analysis

The time_spent field captures when posts were published, but in a relative format rather than absolute timestamps.

Time Value	Count
1 year ago	7,753
2 years ago	5,728
3 years ago	3,759
4 years ago	2,126
3 months ago	1,456
2 months ago	1,448
4 months ago	1,279
10 months ago	1,247
11 months ago	1,133
Other values	~9,000

The relative time format presents a significant limitation for temporal analysis. Without absolute timestamps, we cannot: (1) determine optimal posting times (day of week, hour of day), (2) analyze engagement trends over time, (3) account for LinkedIn algorithm changes, or (4) perform time-series forecasting. The data does show that posts span approximately 4+ years, providing good temporal diversity.

6. Distribution Analysis of Key Metrics

Understanding the distribution of engagement metrics is crucial for several reasons: (1) it informs the choice of statistical methods and machine learning algorithms, (2) it helps identify appropriate data transformations, and (3) it reveals the natural variation in post performance that models must capture.

6.1 Engagement Metrics Distribution

The following visualization shows the distribution of four key metrics: reactions, comments, views, and followers. Due to the extreme skewness of these distributions, a logarithmic transformation (log(x+1)) was applied for visualization purposes.

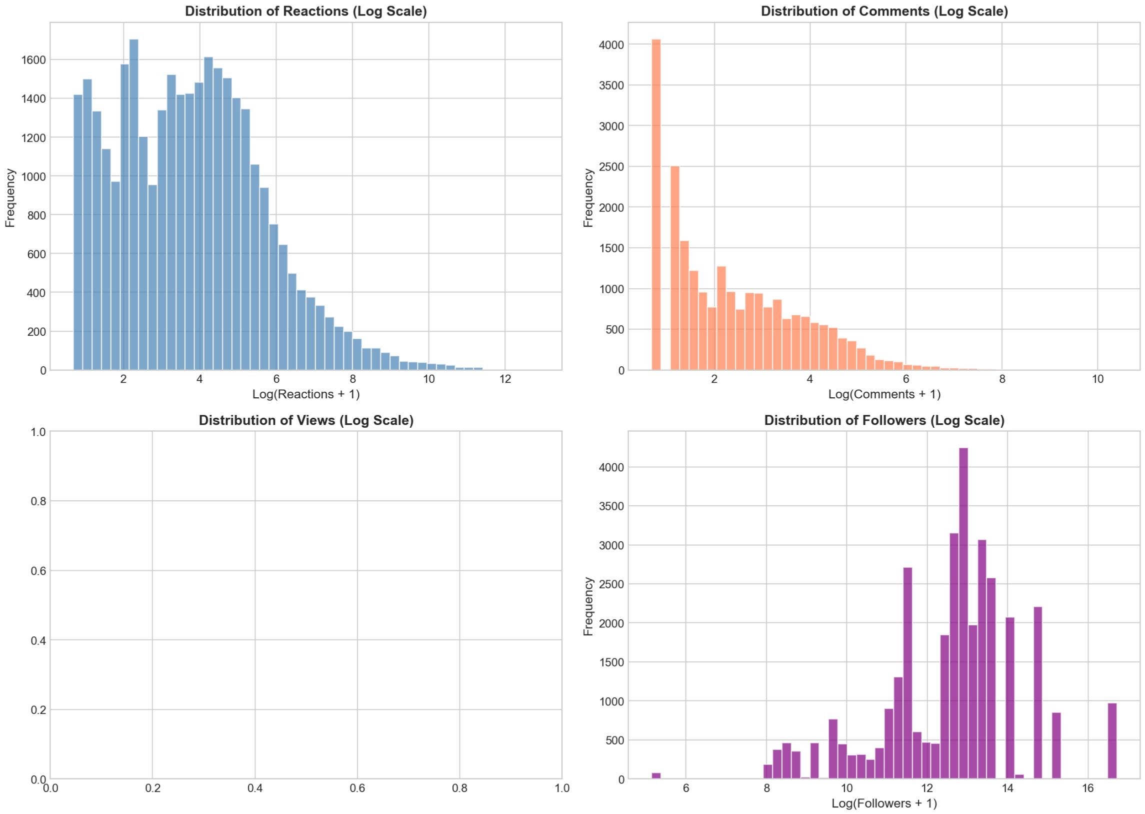


Figure: Distribution of Engagement Metrics (Log Scale)

Interpretation of Engagement Distributions

Reactions Distribution (Top Left): The log-transformed reactions show a roughly normal distribution centered around log(36) ≈ 3.6, indicating that most posts receive between 10-100 reactions. The long right tail represents viral posts with thousands to hundreds of thousands of reactions. This suggests a log-normal distribution is appropriate for modeling.

Comments Distribution (Top Right): Comments follow a similar log-normal pattern but are shifted left, reflecting that comments are less common than reactions. The peak around log(2) ≈ 0.7 indicates that most posts receive 0-5 comments. The 32% of posts with zero comments appear as a spike at the origin.

Views Distribution (Bottom Left): This panel is empty because views data is 100% missing. This represents a critical data gap that prevents reach optimization.

Followers Distribution (Bottom Right): Follower counts span a wide range on the log scale, from around 5 (≈150 followers) to 17 (≈18 million followers). The distribution appears bimodal, suggesting two distinct groups of influencers in the dataset.

6.2 Outlier Detection

Outliers can significantly impact model performance and must be carefully identified and handled. The box plots below show the distribution of each metric on a log scale, highlighting potential outliers as points beyond the whiskers.

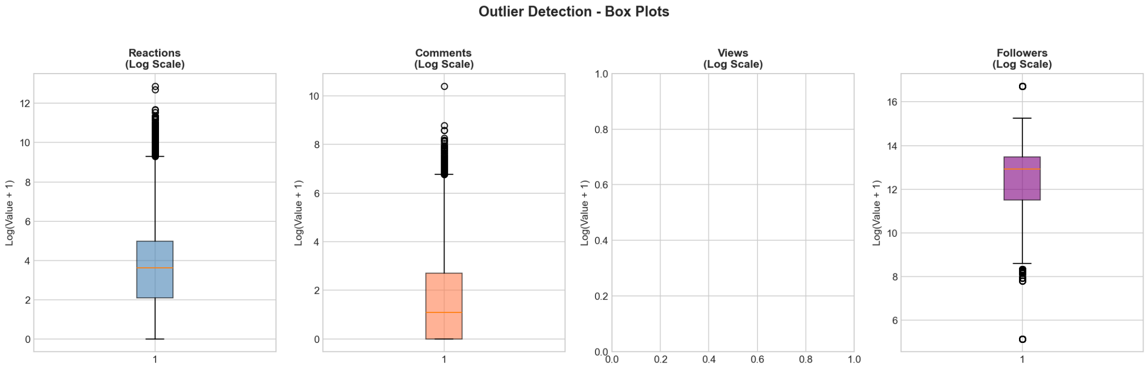


Figure: Outlier Detection Box Plots (Log Scale)

The box plots reveal several important insights about outliers in the dataset:

- All metrics show numerous outliers on the high end, represented by dots above the upper whiskers.
- These outliers represent viral posts that significantly outperform typical content.
- For reactions, the interquartile range (IQR) spans roughly log(7) to log(143), or 7-143 reactions.
- Extreme outliers include posts with 391,498 reactions (max) and 32,907 comments (max).
- For modeling, these outliers could be: (1) capped/winsorized, (2) log-transformed, or (3) modeled separately.

7. Content Analysis

Content characteristics play a crucial role in determining post engagement. This section analyzes the textual properties of posts, including length, word count, hashtag usage, and media type distribution. These insights will inform the content optimization features of the final system.

7.1 Content Length Statistics

Post length can impact readability, engagement, and algorithmic visibility. The following analysis examines both character count and word count distributions.

Metric	Characters	Words
Average	308	49
Median	210	33
Maximum	1,394	260
Empty Posts	2,016 (5.9%)	-

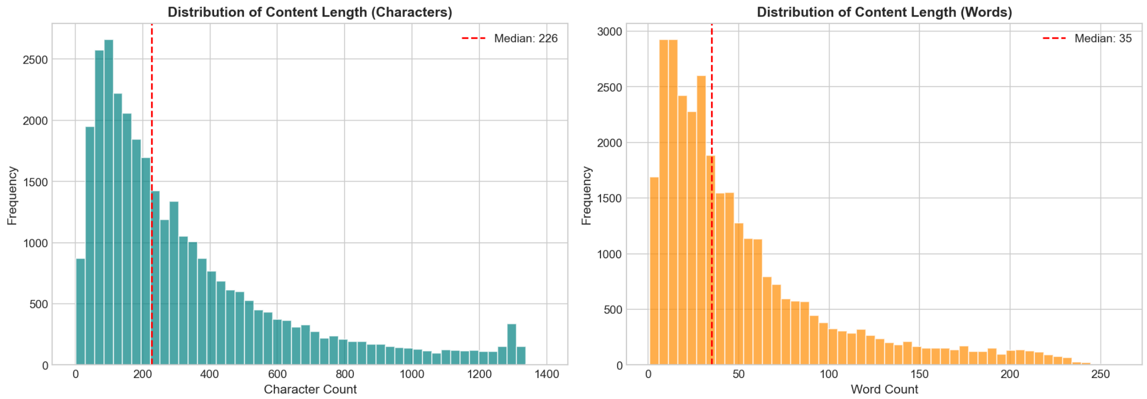


Figure: Distribution of Content Length (Characters and Words)

The content length distributions reveal several patterns relevant to content optimization:

Moderate Length Preference: The median post contains 210 characters (~33 words), suggesting that successful influencers tend to keep posts concise. This aligns with LinkedIn best practices recommending 150-300 characters for optimal engagement.

Right-Skewed Distribution: Both character and word count distributions are right-skewed, with a long tail of longer posts. Some influencers write detailed, longer-form content up to 1,394 characters.

Empty Content: 5.9% of posts have no text content. These are likely media-focused posts (images, videos) that rely on visual content rather than text to convey their message.

7.2 Hashtag Analysis

Hashtags can increase post visibility by connecting content to broader conversations and making posts discoverable to users following specific topics. This analysis examines hashtag usage patterns among influencers.

- **Posts with hashtags:** 14,405 (42.4%)
- **Posts without hashtags:** 19,607 (57.6%)
- **Average hashtags per post:** 2.10
- **Maximum hashtags in a post:** 48
- **Most common count:** 0 (no hashtags)

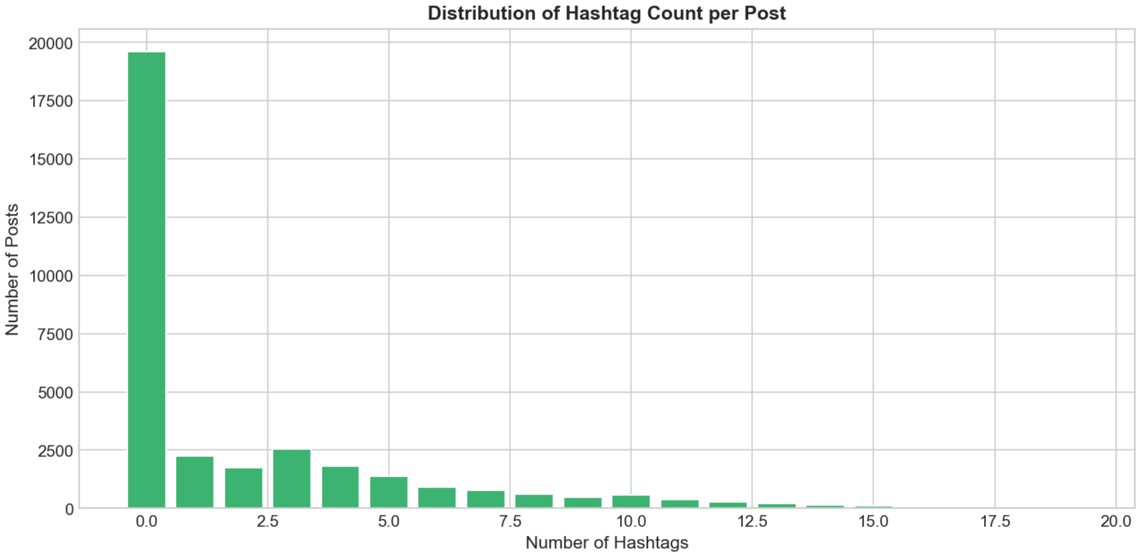


Figure: Distribution of Hashtag Count per Post

The hashtag distribution reveals an interesting pattern: the majority of influencer posts (57.6%) use no hashtags at all. Among posts that do use hashtags, 1-5 hashtags is most common. This challenges the common advice to always include hashtags - successful influencers may rely more on their established audience and content quality than hashtag-based discovery. However, the correlation between hashtag count and engagement should be examined before drawing conclusions.

7.3 Media Type Distribution

The type of media attached to a post can significantly impact its performance. LinkedIn supports various media types including articles, images, videos, documents, and polls.

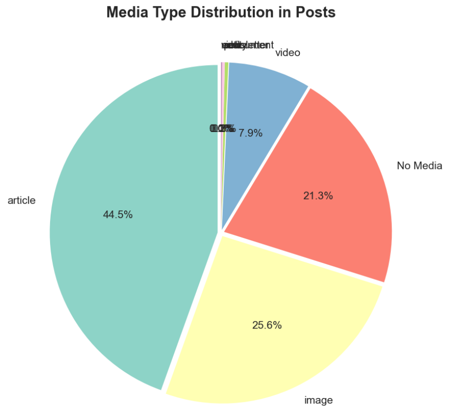


Figure: Distribution of Media Types in Posts

The pie chart reveals the following media type distribution:

Articles (44.5%): Nearly half of all posts include article links, reflecting LinkedIn's positioning as a platform for sharing professional insights and external content.

Images (25.6%): About a quarter of posts feature images, used for infographics, quotes, and visual storytelling.

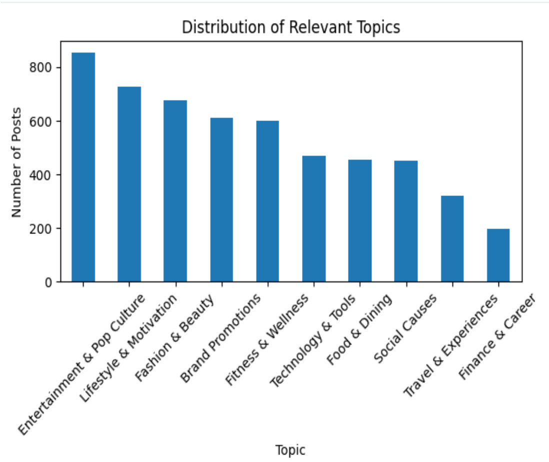
No Media (21.3%): Over one-fifth of posts are text-only, demonstrating that compelling text content can stand on its own without media attachments.

Videos (7.9%): Video content represents a smaller but significant portion of posts. As we'll see in the engagement analysis, videos often achieve the highest engagement despite their lower frequency.

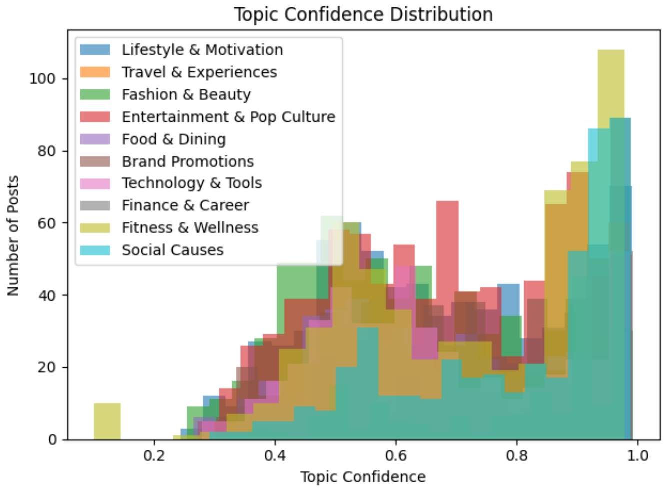
Other (0.7%): Documents, polls, newsletters, and other media types make up a small fraction of posts.

7.4 Topic Modelling

Topic modeling was performed on a subset of the dataset consisting of posts from the most recent three months, rather than the full historical data. This design choice was intentional, as influencer content trends—particularly on professional platforms like LinkedIn—are highly time-sensitive. Restricting the analysis to recent data ensures that the extracted topics reflect current audience interests, platform norms, and emerging trends, rather than outdated or no-longer-relevant themes.



The topic distribution reveals that **Entertainment & Pop Culture** and **Lifestyle & Motivation** dominate recent influencer content, indicating a strong emphasis on broad, engagement-driven themes. In contrast, **Finance & Career** and **Travel & Experiences** are comparatively underrepresented, suggesting that these topics occupy more specialized niches. **Technology & Tools** and **Brand Promotions** show moderate representation, highlighting a growing presence of professional and commercial discourse within influencer content.



The topic confidence distribution demonstrates that most topics achieve **high confidence scores (primarily between 0.6 and 0.9)**, indicating well-defined and semantically coherent topic clusters. Topics such as **Fitness & Wellness**, **Social Causes**, and **Lifestyle & Motivation** exhibit particularly strong confidence, reflecting consistent language usage. Lower confidence dispersion observed in **Brand Promotions** and **Entertainment & Pop Culture** suggests mixed or hybrid content styles, where promotional messaging or storytelling spans multiple themes.

Overall, using a recent data subset improves the **relevance, interpretability, and actionability** of the topic modeling results. The findings indicate that while general-interest content dominates in volume, professionally oriented topics especially **Finance & Career** and **Technology & Tools** are less frequent yet semantically robust, making them well-suited for targeted LinkedIn content strategies and thought leadership initiatives.

8. Engagement Analysis

Engagement metrics are the primary target variables for the TrendPilot system. Understanding how engagement varies across different content types and influencers is essential for building effective recommendation models. This section provides detailed analysis of engagement patterns.

8.1 Engagement by Media Type

Different media types generate significantly different levels of engagement. The following analysis compares average and median engagement metrics across all media types in the dataset.

Media Type	Count	Avg Reactions	Med. Reactions	Avg Comments	Med. Comments
Video	2,690	866	114	70	12.5
Image	8,708	824	87.5	40	6
Article	15,144	161	13	10	1
Document	113	95	51	29	9
Poll	86	58	29.5	34	22.5
Entity	32	50	13.5	4	0
Newsletter	4	17.5	19	1	1
View	2	4	4	0	0

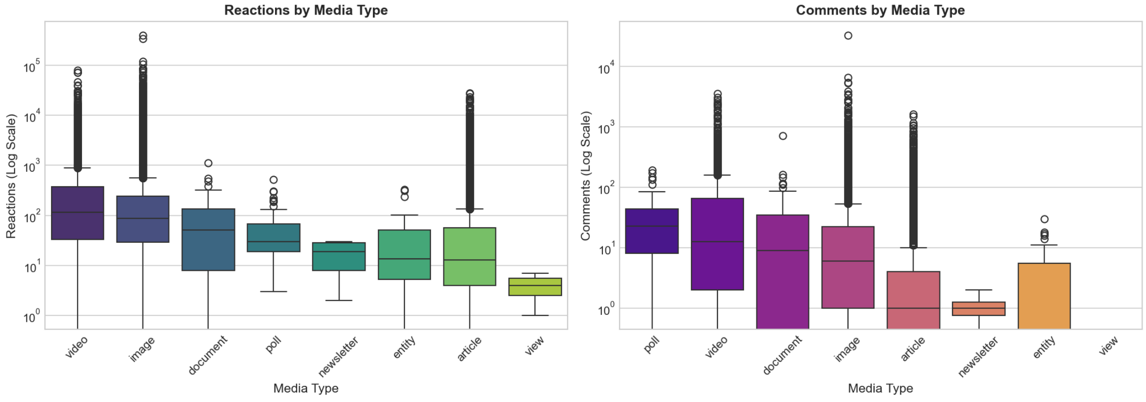


Figure: Engagement Comparison by Media Type (Log Scale)

Key Insights from Media Type Analysis

Video Dominates Engagement: Video content generates the highest average engagement with 866 reactions and 70 comments per post. The median video post receives 114 reactions - nearly 9x the median article. This strongly suggests that users should prioritize video content when seeking maximum engagement.

Images Outperform Articles: Image posts average 824 reactions compared to 161 for articles. Despite articles being the most common media type (44.5% of posts), they generate the lowest engagement per post among the major categories. This represents a potential optimization opportunity.

Polls Drive Comments: While polls have lower reaction counts, they generate disproportionately high comment engagement (median 22.5 comments). Polls may be effective for starting conversations and increasing audience interaction.

Document Performance: Documents (PDFs, slideshows) show solid performance with median 51 reactions, suggesting they're effective for sharing detailed professional content.

8.2 Engagement Rate Analysis

Raw engagement counts can be misleading because influencers with larger followings naturally receive more engagement. Engagement rate normalizes for audience size, providing a fairer comparison of content effectiveness. We calculate engagement rate as reactions per 1,000 followers.

Metric	Value
Mean Engagement Rate	0.77 reactions per 1K followers
Median Engagement Rate	0.11 reactions per 1K followers
Standard Deviation	3.91

The large gap between mean (0.77) and median (0.11) engagement rates indicates high variability in content performance. Even among established influencers, most posts achieve modest engagement rates, while a small percentage of posts dramatically outperform the rest. This suggests that content quality and relevance matter more than follower count alone.

9. Correlation Analysis

Understanding the relationships between different features is essential for feature engineering and identifying predictors of engagement. This section examines correlations between key metrics using Pearson correlation coefficients.

9.1 Correlation Matrix

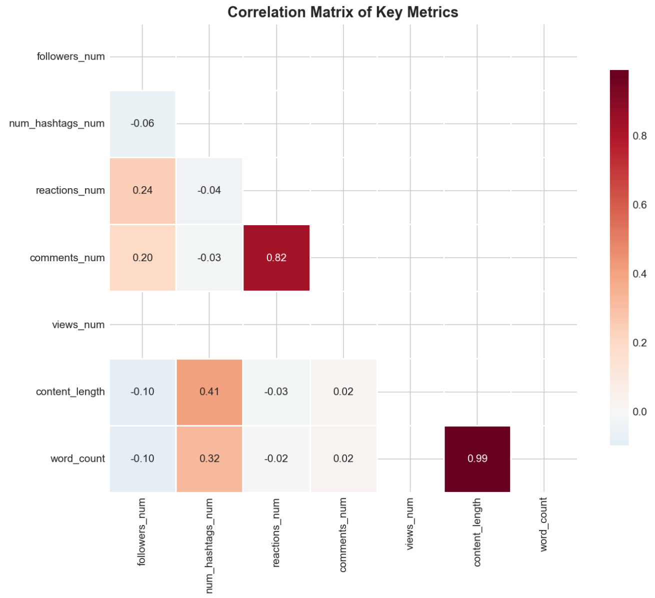


Figure: Correlation Matrix of Key Metrics

9.2 Key Correlation Findings

Strong Correlations

Reactions and Comments (r = 0.823): The strongest correlation in the dataset exists between reactions and comments. This makes intuitive sense - posts that attract reactions also tend to generate discussion. This high correlation suggests that these two metrics could potentially be combined into a single "engagement score" for modeling purposes, reducing dimensionality while preserving signal.

Moderate Correlations

Followers vs Reactions (r = 0.242): A weak positive correlation exists between follower count and reactions. Larger audiences naturally lead to more reactions, but the moderate strength suggests content quality matters more than audience size.

Followers vs Comments (r = 0.195): Similar to reactions, comments show a weak positive correlation with followers. The slightly lower coefficient suggests that comments depend more on content that prompts discussion rather than raw audience size.

Negligible Correlations

Several features show negligible correlation with engagement, which is itself an important finding:

Num Hashtags vs Reactions (r = -0.042): Contrary to popular belief, hashtag count shows almost no correlation with reactions. In fact, the very slight negative correlation suggests that excessive hashtag use might even slightly hurt engagement.

Content Length vs Reactions (r = -0.027): Post length has virtually no impact on engagement, suggesting that both short and long posts can be equally successful if the content is compelling.

Word Count vs Comments (r = 0.025): Similarly, word count shows no meaningful correlation with comments, indicating that prompting discussion depends on content substance rather than length.

9.3 Implications for Modeling

The correlation analysis has several important implications for building engagement prediction models:

- Reactions and comments can potentially be combined into a single target variable due to their strong correlation.
- Follower count should be included as a feature but may benefit from interaction terms or normalization.
- Hashtag count and content length are unlikely to be strong predictors on their own.
- Media type (categorical) and content semantics (from NLP) may be more important predictors than simple numerical features.
- Feature engineering should focus on content quality metrics rather than quantity metrics.

10. Top Performers Analysis

Analyzing the top-performing influencers provides insights into what success looks like on LinkedIn and helps validate that the data captures meaningful variation in performance. This section examines the highest-engagement influencers and their posting patterns.

10.1 Top Influencers by Average Engagement

The following table shows the top 10 influencers ranked by average reactions per post, filtered to include only those with at least 5 posts to ensure statistical reliability.

Rank	Name	Avg Reactions	Total Posts	Followers	Avg Comments
1	Simon Sinek	16,641	264	4.2M	441
2	Richard Branson	5,186	975	18.3M	236
3	Kevin O'Leary	2,784	424	2.8M	193
4	Ian Bremmer	2,158	590	3.7M	99
5	Vani Kola	1,166	949	1.2M	46
6	Quentin M. Allums	490	366	66K	119

7	Tom Goodwin	375	1,272	719K	67
8	James Altucher	338	1,125	1.3M	27
9	Natalie Riso	248	195	406K	22
10	Tai T.	242	251	348K	21

10.2 Observations from Top Performers

Exceptional Performance: Simon Sinek stands out dramatically with an average of 16,641 reactions per post - over 3x the second-place Richard Branson. His content consistently resonates with his audience, making his posts valuable case studies for content optimization.

Follower Count Not Deterministic: Richard Branson has the most followers (18.3M) but ranks second in engagement. Meanwhile, Quentin M. Allums achieves high engagement (490 avg reactions) with only 66K followers, demonstrating that audience engagement depends on more than audience size.

Posting Frequency Varies: Top performers show varied posting frequencies from 195 posts (Natalie Riso) to 1,272 posts (Tom Goodwin). There's no clear correlation between posting volume and average engagement, suggesting quality matters more than quantity.

Comments vs Reactions Ratio: Simon Sinek also leads in comments with 441 average comments per post, but the reactions-to-comments ratio varies across influencers. Some generate more discussion relative to their reactions, indicating different content styles.

11. Data Gap Risk Assessment

This section provides a comprehensive risk assessment of identified data gaps, categorizing each by severity and analyzing the specific impact on planned system features. Understanding these risks is essential for project planning and setting appropriate expectations.

11.1 Risk Assessment Matrix

Data Gap	**Missing %**	**Severity**	**Impact on Project**
Views data	100%	HIGH	Cannot optimize for reach/virality
Absolute timestamps	100%*	HIGH	Cannot recommend posting times
Hashtag followers	100%**	MEDIUM	Cannot assess hashtag potential
Connections	24.4%	LOW	Field has limited utility anyway
Media type	21.3%	MEDIUM	Reduced sample for media analysis

Location	6.7%	MEDIUM	Limited geographic targeting
Content	5.9%	LOW	Minimal impact - 94% available
Followers	0.12%	LOW	Easily imputable

*Timestamps exist but in relative format only (e.g., "2 days ago")

**Column exists but all values are 0

11.2 Detailed Risk Analysis

HIGH Severity Risks

Views Data Gap

The complete absence of views data represents the most significant data gap. Views/impressions are a fundamental metric for understanding post reach and calculating conversion rates (views-to-engagement). Without this data, we cannot: (1) distinguish between posts that reached many people but received low engagement vs posts with limited reach, (2) optimize content for visibility in the LinkedIn feed, (3) calculate impression-based engagement rates for fairer comparisons.

MITIGATION: Focus on engagement optimization rather than reach optimization. Use follower count as a proxy for potential reach. Consider this a limitation in the final system documentation.

Timestamp Format Gap

While time data exists, it's in relative format ("1 week ago") rather than absolute timestamps. This prevents temporal analysis including: (1) optimal posting day of week, (2) optimal posting hour, (3) engagement decay patterns over time, (4) seasonal or trending topic analysis.

MITIGATION: Acknowledge this limitation. For future data collection, prioritize capturing absolute timestamps. Consider using LinkedIn's API for time-sensitive analysis.

MEDIUM Severity Risks

Media Type Gap (21.3%)

About one-fifth of posts lack media type classification. This reduces the sample size for media-based analysis and may introduce bias if certain media types are more likely to have missing values.

MITIGATION: Analyze patterns in missing media types. Consider treating "No Media" as its own category. Ensure models handle missing values appropriately.

Location Gap (6.7%)
Geographic data is missing for some records, limiting geographic targeting features.

MITIGATION: Geographic features may not be critical for content optimization. Consider deprioritizing location-based recommendations.

12. Impact Analysis on Final Project

The TrendPilot system aims to help users create engaging LinkedIn posts. This section evaluates the feasibility and scope of each planned feature based on the data availability and quality findings from this EDA.

12.1 Feature Feasibility Assessment

Content Optimization Features

Data Available	94.1% of posts have content
Status	FULLY FEASIBLE
Confidence	HIGH

Capabilities enabled by available data:

- Analyze text patterns, structure, and writing style of high-engagement posts
- Build NLP models to suggest content improvements
- Identify keywords, phrases, and topics associated with high engagement
- Recommend optimal content length based on engagement correlations
- Detect sentiment and tone patterns in successful posts

Hashtag Recommendation Features

Data Available	100% hashtag lists, 0% follower counts
Status	PARTIALLY FEASIBLE
Confidence	MEDIUM

Available capabilities:

- Recommend hashtags based on content topic matching
- Analyze which hashtags correlate with higher engagement

- Suggest optimal number of hashtags based on data patterns

Unavailable due to data gaps:

- Recommend hashtags by reach potential (no follower data)
- Predict hashtag-driven visibility improvements

Media Type Recommendations

Data Available	78.7% of posts have media type
Status	FULLY FEASIBLE
Confidence	HIGH

Capabilities enabled by available data:

- Recommend optimal media type based on content and goals
- Quantify engagement differences between media types
- Suggest video/image when higher engagement is the goal
- Identify when text-only posts are appropriate

Engagement Prediction Model

ML-Ready Rows	26,107 (76.8%)
Status	FULLY FEASIBLE
Confidence	HIGH

Model capabilities:

- Predict expected reactions based on content features
- Predict expected comments based on content features
- Provide confidence intervals for predictions
- Identify factors most predictive of engagement

Critical limitation:

- Cannot predict or optimize for views/reach due to 100% missing views data

Posting Time Recommendations

Data Available	Relative time only
Status	NOT FEASIBLE
Confidence	N/A

This feature cannot be implemented with the current data. The relative timestamp format ("2 weeks ago") does not provide the day-of-week or hour-of-day information needed for posting time optimization. This feature should be deprioritized or planned for a future phase with new data collection.

13. Recommendations

Based on the findings of this EDA, the following recommendations are provided for the development of the TrendPilot system and for future data collection efforts.

13.1 Feature Prioritization

Tier 1: Focus Features (Sufficient Data)

These features should be the primary focus of development:

Content Text Analysis: Build NLP models to analyze content patterns, suggest improvements, and predict engagement.

Media Type Recommendations: Recommend optimal media types based on clear engagement differences in the data.

Engagement Prediction: Develop ML models to predict reactions and comments based on content features.

Hashtag Suggestions: Recommend relevant hashtags based on content, even without reach data.

Tier 2: Limited Features (Partial Data)

These features can be implemented with acknowledged limitations:

Engagement Rate Normalization: Use follower count to normalize engagement, acknowledging some missing values.

Content Length Guidance: Provide soft recommendations based on observed patterns, though correlation is weak.

Tier 3: Deprioritized Features (Insufficient Data)

These features should be deprioritized or excluded:

Views/Reach Optimization: Cannot implement - views data is 100% missing.

Posting Time Recommendations: Cannot implement - no absolute timestamps available.

Geographic Targeting: Limited value - location data is sparse and may not impact content strategy.

Hashtag Reach Prediction: Cannot implement - hashtag follower data is all zeros.

13.2 Data Collection Recommendations

For future data collection phases, prioritize the following:

Absolute Timestamps: Capture the exact date and time of each post to enable temporal analysis and posting time recommendations.

Views/Impressions: Ensure views data is captured consistently. This may require different collection methods or API access.

Hashtag Metadata: Collect hashtag follower counts when available to enable reach-based recommendations.

Engagement Over Time: Track how engagement accumulates over time (1 hour, 24 hours, 1 week) to understand decay patterns.

Profile Updates: Capture follower growth over time to understand audience building patterns.

13.3 Model Development Recommendations

Primary Target Variable: Use reactions + comments as the primary engagement metric. Given their strong correlation (r=0.823), consider combining them into a single score or predicting them jointly.

Handle Skewness: Apply log transformation to engagement metrics before modeling. The extreme skewness of raw values will otherwise dominate model training.

Engagement Rate: Consider using engagement rate (reactions/followers) as an alternative target to account for audience size differences.

Feature Engineering: Focus on content-derived features (sentiment, topics, readability) rather than simple counts. Correlation analysis shows counts have limited predictive power.

Missing Value Strategy: For media_type, treat missing as a "No Media" category. For other features, use appropriate imputation or exclusion depending on the modeling approach.

Validation Strategy: Use influencer-stratified cross-validation to ensure models generalize across different posting styles.

14. Conclusion

14.1 Summary of Findings

This exploratory data analysis has provided a comprehensive assessment of the LinkedIn influencer dataset for the TrendPilot project. The dataset comprises 34,012 posts from 69 influential LinkedIn users, offering a substantial foundation for building engagement prediction and content optimization models.

The analysis revealed that the dataset achieves strong data quality for core features, with 95.4% completeness for the features most critical to engagement prediction (content, reactions, comments, media type, hashtags, and followers). This level of completeness is sufficient for developing reliable machine learning models.

However, significant gaps exist that limit certain capabilities. Most notably, views data is completely missing (100%), preventing reach optimization. Additionally, timestamps are stored in relative format, preventing temporal analysis and posting time recommendations. These limitations should be clearly communicated in the final system and addressed in future data collection efforts.

14.2 Data Quality Score

Overall Data Completeness	86.7%
Core Features Completeness	95.4%
Quality Assessment	GOOD
Suitability for ML	SUITABLE

14.3 Key Takeaways

- The dataset is suitable for building a LinkedIn post optimization system focused on content analysis, media type recommendations, and engagement prediction.
- Video and image content significantly outperform articles in engagement, suggesting clear recommendations for users seeking maximum impact.
- Reactions and comments are strongly correlated (r=0.823), enabling their combination into a unified engagement metric.
- Hashtag count and content length show negligible correlation with engagement, challenging conventional wisdom.
- Top performers like Simon Sinek achieve exceptional engagement through content quality, not just audience size.
- Future data collection should prioritize absolute timestamps and views data to enable currently infeasible features.

14.4 Next Steps

1. Proceed with feature engineering based on EDA insights

3. Build engagement prediction models using available features

4. Create media type recommendation logic

5. Document limitations for user transparency

6. Plan future data collection to address identified gaps

This EDA provides a solid foundation for the development of TrendPilot. While some features will be limited by data gaps, the available data supports the core goal of helping users create more engaging LinkedIn content through data-driven recommendations.