TRANSFER PROJECT

Hochschule Luzern – Informatik CAS Machine Learning

Success Predictor for LinkedIn Posts

In general vs. Before vs. After the LinkedIn algorithm changes in 2025.

Author:
Reto Lämmler
Rotachstrasse 30
8003 Zürich

Date of Submission: 1.2.2026

Programme Director: Prof. Dr. Umberto Michelucci

Mentor: Aygul Zagidullina

Confidentiality: Non-confidential

Affidavit

I hereby declare that I have written this paper independently and without the assistance of third parties. I have used only the sources and resources indicated, with full disclosure of the generative AI systems used to complete this project:

- **Code Development:** I used Large Language Models (LLMs) to generate code and figures utilized in this project, specifically Claude, Gemini, and GitHub Co-Pilot.
- **Documentation:** I used Large Language Models (LLMs) to structure, polish and optimize the language of this documentation, specifically ChatGPT, Claude and Gemini.

I have identified all ideas taken verbatim or in substance from sources and provided the necessary source references, including illustrations and tables. I will respect the copyright provisions of Lucerne University of Applied Sciences and Arts.

Reto Lämmler

Intellectual property in accordance with the study regulations for education at Lucerne University of Applied Sciences and Arts, FH Central Switzerland.

Abstract

The LinkedIn algorithm has shifted from a social-graph-based system to a retrieval-based model using large language models to match users with relevant content. This shift suggests that content relevance and creator-related signals (e.g., network size and topic authority) are becoming more important than simple engagement rules (e.g., posting time and image). This project analyzes which factors best predict LinkedIn post performance by combining structural features such as content length, media presence, formatting, network size, and posting time (hour of day and day of week) with NLP-based features including semantic alignment and sentiment.

The study uses a dataset of around 10,000 LinkedIn posts from 42 different users. To avoid identity bias, several preprocessing techniques were applied. These included limiting the number of posts per user, balancing high- and low-performing posts within each user, and reducing the influence of extreme viral outliers using capping and a relative engagement score. An optimized XGBoost binary classification model was trained to predict whether a post would perform above or below a user's typical engagement level.

The final model achieved an accuracy of 67%, representing a 1% improvement over a Random Forest base model and 17% improvement over random classification, with balanced performance across both classes. SHAP analysis showed that images, post content length, and network size were the strongest drivers of engagement. Semantic alignment between a user's profile and post content had a positive but smaller effect. External links mostly reduced post performance, while posting time variables (hour of day and day of week) showed minimal influence on engagement outcomes.

A preliminary time-based SHAP comparison explored potential concept drift following LinkedIn's reported algorithm changes in early 2025. The analysis suggests a possible shift away from the existing success drivers towards greater emphasis on content depth, creator authority, and semantic relevance. However, due to the limited size of the post–algorithm-change dataset, these temporal findings should be considered directional hypotheses rather than definitive conclusions.

TABLE OF CONTENTS

1 Introduction

1.1 Background and Motivation

1.1.1 The Strategic Role of LinkedIn in Professional Communication

LinkedIn has transitioned from a digital resume repository into a central platform for B2B marketing, personal branding, and thought leadership. With over one billion users , the platform now functions as a "content-first" ecosystem where professional success is increasingly tied to one's ability to get exposure in the social feed. For professionals and businesses alike, a high-performing post is no longer just a nice to have; it is a driver of authority, lead generation, and career opportunities.

1.1.2 The ever changing LinkedIn algorithm

The mechanism that determines which content goes viral and which one does not is proprietary knowledge and only partially shared by LinkedIn. It’s an ongoing topic discussed by the LinkedIn community and marketing experts, often inflated with assumptions and myths. Content creators often discuss "hacks" or trial-and-error strategies that fail to reliably leverage the platform’s ongoing algorithm changes.

Recent algorithm changes in 2025 have further complicated this landscape, as the algorithm has evolved to prioritize the creator’s authority and meaningful engagement over simple likes or click bait like “Comment YES, if you agree!”. Furthermore, the platform has introduced more aggressive filters for AI-generated content.

1.1.3 Motivation: From Speculation to Algorithmic Transparency

To move beyond speculative writing advice, this research uses an XGBoost classifier and SHAP values to decode LinkedIn post performance. While the model distinguishes successful from unsuccessful content, SHAP provides the necessary transparency by quantifying the impact of features like post content size, structure, semantic alignment etc.

1.1.4 Concept Drift: LinkedIn Algorithm change in 2025

Finally, this transfer project looks also into the ongoing Concept Drift, the reality that social media algorithms are not static. A strategy that worked in 2024 may result in a penalty in 2026. By analyzing the shifts in feature importance by splitting the data set to analyze the feature importance in regards to the most recent algorithm change in 2025.

1.2 Problem Statement

There is limited empirical evidence on which features currently drive engagement on LinkedIn. In particular, it remains unclear whether the creator's authority, content quality or timing is the strongest predictor for post performance. Additionally, observations suggest that a potential concept drift due to an algorithm change in 2025 could have altered the relative importance of these features.

1.3 Objectives and Research Question

The primary objective of this project is to build a machine-learning model that identifies the strongest predictors of LinkedIn engagement to predict whether a future post will be high or low performing (binary classification). One of the central research questions examines whether semantic alignment (vector distance between a LinkedIn profile to post content) and post timing are dominant drivers of engagement and how they changed in regards to the 2025 algorithm change.

1.4 Scope and limitations

The analysis is based on approximately 10'000 LinkedIn posts voluntarily provided by 42 LinkedIn users. After cleanup and balancing, approximately half (5'000 LinkedIn posts) could be used for the model training. The scope is limited to content size, structure, timing and network size. The post content itself was assessed for semantic alignment with the user profile as well as the sentiment. Image and video content are included only as binary indicators (Has Image, Has Video), while computer-vision analysis of the actual media is excluded.

2 Literature Review

2.1 Expert Perspectives and Industry Insights

Recent industry analysis reveals a fundamental shift in the LinkedIn algorithm from engagement-based virality to expertise-driven distribution. This section summarizes key findings from Hootsuite (Newberry & Christison, n.d.) and practitioner Chad Johnson (Johnson, 2025) regarding the 2025 updates.

2.1.1 Algorithmic Filtering and Distribution

The LinkedIn feed operates on a multi-stage filtering process designed to prioritize professional authority. Initially, content is categorized as spam, low quality, or high quality based on structural signals. High-quality content enters a "testing phase" where it is shown to a small audience sample to measure initial relevance.

Expert consensus emphasizes that dwell time (driven by compelling hooks in the first couple of lines) is now a dominant ranking signal. Furthermore, the platform maintains a "walled garden" approach, where native content is prioritized over posts containing external links, which often face significant reach penalties.

2.1.2 The "360 Brew" System and Semantic Authority

Chad Johnson discusses the introduction of the 360 Brew model, a large-scale language model, that has shifted the platform toward deep contextual matching. This system evaluates semantic alignment, measuring how closely a post relates to the author’s established professional profile. Content that deviates from an author's proven niche is frequently suppressed.

This model has also redefined engagement hierarchies. Simple likes have been deprioritized in favor of saves and long-form comments. Multi-sentence responses are treated as high-value signals of professional discourse, prompting the algorithm to distribute the content to a wider, relevant audience.

2.1.3 Synthesis for Model Development

These expert insights provide the qualitative basis for this project’s feature engineering. The algorithmic shift toward post quality (hooks, length), the growing emphasis on semantic alignment, and the platform's bias against external redirects directly informed the selection and importance of features like Link Count, Hook Length, and Semantic Alignment in the subsequent XGBoost analysis.

2.2 Large Scale Retrieval and Generative Architectures

LinkedIn feed has transitioned from traditional heuristic ranking to a sophisticated retrieval system based on generative artificial intelligence. Recent research (Ramanujam et al., 2025) details the implementation of a generative retrieval and ranking framework that utilizes Causal Language Models such as LLaMA 3 to improve content relevance and member engagement.

2.2.1 The Two Tower Architecture

At the core of this system is a Two Tower architecture designed for extreme scale. In this framework, the retrieval process is split into two distinct components:

- **The Member Tower:** This component encodes user data, including professional history, interests, and recent interactions, into a dense vector embedding.
- **The Item Tower:** This component simultaneously encodes the content of a post, including text, media type, and metadata, into a corresponding vector space.

Success is predicted by calculating the cosine similarity between these two embeddings. A higher similarity score indicates a greater probability of engagement, as it suggests the content aligns closely with the professional identity and consumption habits of the user.

2.2.2 Theoretical Justification for Semantic Alignment

This shift provides the foundation for the Semantic Alignment feature engineered in this project. By calculating the cosine similarity between a user profile summary and their post content, it simulates the internal retrieval logic used by LinkedIn.

While classic features like posting time focus on the logistics of distribution, Semantic Alignment focuses on the thematic relevance that current retrieval models prioritize. The inclusion of this feature allows the XGBoost model to test whether being a "topical authority" as defined by the Two Tower system is indeed a primary driver of performance in the current algorithmic era.

3 Data Acquisition

The data acquisition phase required a strategic approach to balance data quality with technical and legal constraints. Several alternatives for retrieving detailed LinkedIn user data were evaluated.

3.1 Evaluation of Data Retrieval Methods

Retrieving the required data from LinkedIn presented several significant challenges:

- **LinkedIn API Constraints:** Accessing detailed post performance and member data via the official LinkedIn API requires "Partner Access." For a transfer project of this scope, obtaining such permission was not a viable option.
- **Scraped Datasets:** Brightdata (brightdata, -) sells scraped LinkedIn data incl. posts. Besides the “grey zone”, a trial access revealed low data quality and the lack of critical metadata required for feature engineering.
- **User Data Donation:** In light of these limitations, the only viable path to acquiring a high-fidelity, ethical, and accurate dataset was to reach out to the author’s professional network and request a "data donation." This approach ensured that the data was obtained with explicit consent and directly from the platform’s own records.

3.2 LinkedIn Data Archive Export

Data was acquired via the LinkedIn Data Archive export. This method was selected to ensure data integrity, as it provides an archive with high-fidelity CSV files. Only the following core files were needed for this project:

- **Shares.csv:** Contains the historical record of all user posts, including raw timestamps and the full text of the content. The file didn't contain post performance data like likes and comments.
- **Connections.csv:** Provides a chronological record of network growth, later used to estimate network size at the time of each post.
- **Profile.csv:** Includes professional metadata such as headlines and bios, which are essential for calculating the semantic alignment between the author and their content.

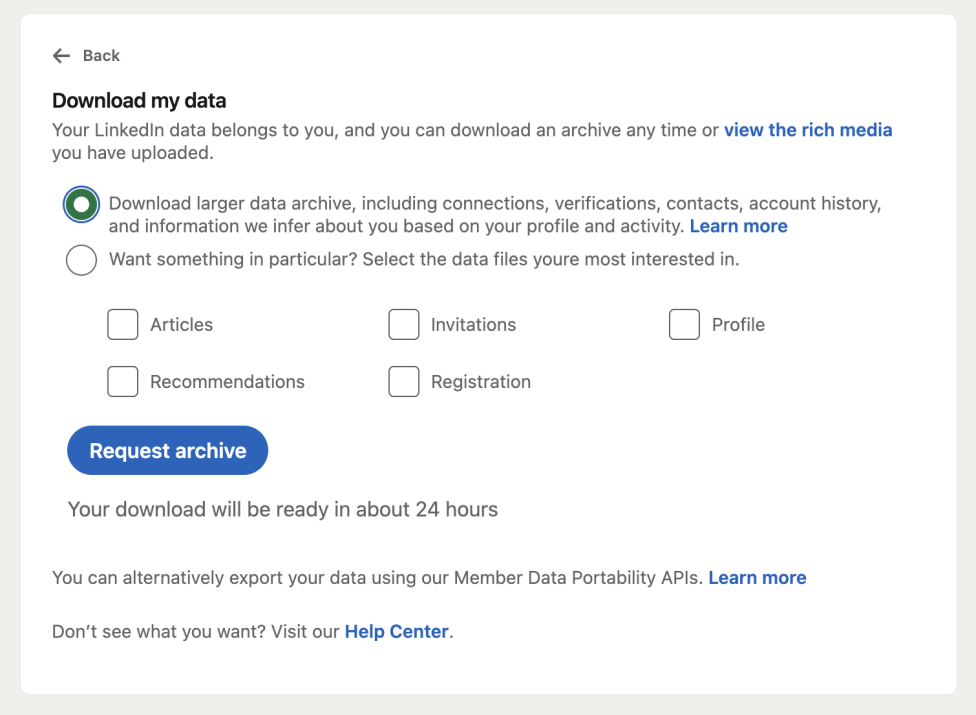

Figure 1 - Screenshot Download my data on LinkedIn

3.3 LinkedIn User Outreach and Participant Cohort

To build a model capable of generalizing beyond a single creator style, data was aggregated from a diverse cohort of 42 participants. This group ranged from casual users with as few as seven posts to power users with over 1,000 posts. While more than 50 individuals originally provided their data, the LinkedIn export failed to function correctly for approximately 10 percent of the participants. The most common issue was the absence of the Shares.csv file in the provided archives.

The outreach was conducted through a LinkedIn post (Lämmler, 2025) designed to trigger the very engagement mechanics this study investigates. The post utilized a "comment-to-receive" mechanism, asking interested participants to comment “DATA” to receive sharing instructions and a link to a Google Forms survey. This survey served as the secure intake point for the CSV files and allowed participants to opt-in to the study.

Although the retrieved dataset is subject to requester bias since it was sourced through the author's personal network, the diversity within the cohort provides a strong base for training. This diversity allows the final XGBoost model to identify structural and semantic success patterns that generalize across a broad spectrum of creators.

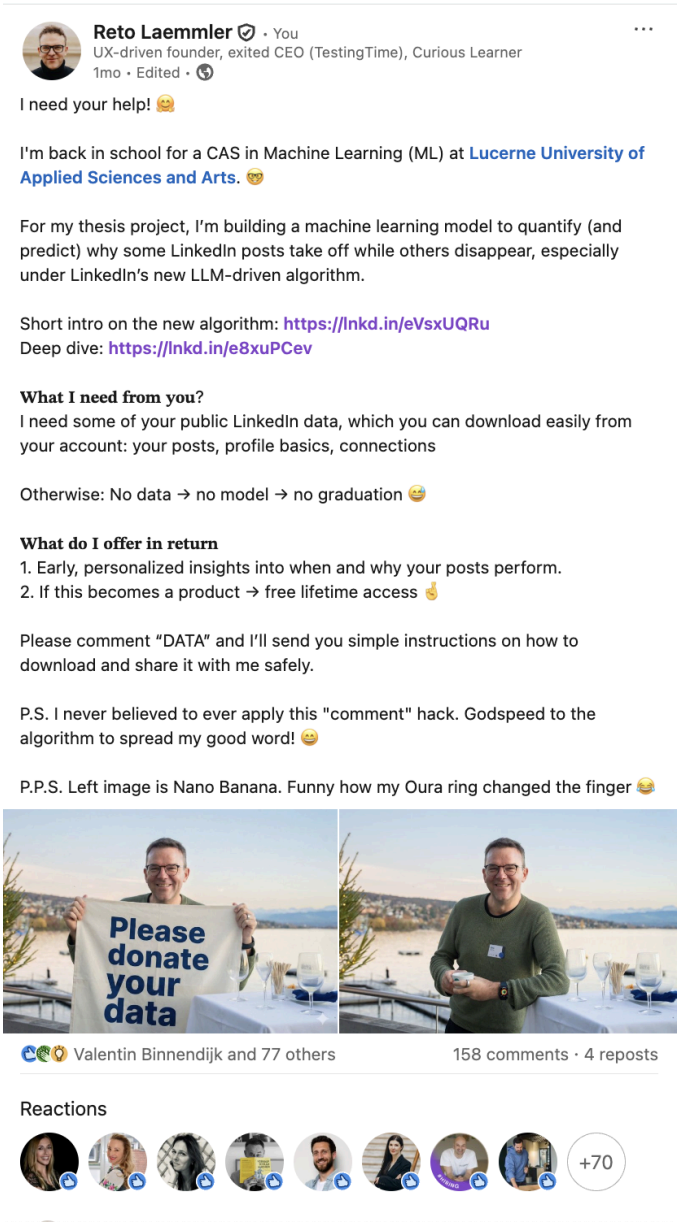

Figure 2 - Screenshot LinkedIn Post

4 Data Preprocessing

The raw dataset contained 10912 unique posts but lacked the engagement and performance metrics and context required for sophisticated modeling.

4.1 Data Enrichment and Consolidation

A custom pipeline was developed to enrich and consolidate the data:

- **Engagement and Media Enrichment:** Because the standard archive lacks post performance data, a specialized post extractor tool (Fox, 2023) was utilized to retrieve metadata including the presence of images, videos, or documents, as well as the total count of comments and likes. It should be noted that the number of reposts and impressions was not extractable and is therefore excluded from the analysis.
- **Network Size at Posting Time:** To calculate a relative engagement score for each post, it was necessary to estimate the network size of a user at the specific moment of a post publication. Using a `pd.merge_asof` backward search strategy, the Shares.csv was consolidated with the Connections.csv. This allows the calculation of an exact network size for every historical post by matching the post timestamp with the nearest preceding connection date.
- **Profile Integration:** From the Profile.csv, the headline and summary were extracted and merged into the dataset. This step was critical for enabling subsequent semantic analysis to measure topical alignment between profile and post content.

4.2 Data Cleanup and Quality Assurance

Following cleanup steps were applied, which resulted in a final dataset of 8542 posts long:

- **Deduplication and Noise Filtering:** Duplicate URLs and redundant entries were removed to prevent biases. Furthermore, posts with empty content were filtered out since they were of no use for the calculation of the semantic alignment nor sentiment. This exclusion was critical as "low-context" posts often represent reshares without any original content.
- **Robust Timestamp Parsing:** LinkedIn exports occasionally exhibit inconsistent date formats depending on the region or version of the archive. A multi-format parsing strategy was implemented to standardize all timestamps into a unified temporal coordinate system, ensuring accurate longitudinal analysis.

After all the data preprocessing steps, the dataset contained following fields:

- **User ID:** Firstname + Lastname from Profile.csv
- **Profile Summary:** Headline + Summary from Profile.csv
- **Post URL:** Link to the post from Shares.csv

- **Post Timestamp DT:** Timestamp in GMT+1from Shares.csv
- **Post Content:** Raw post content from Shares.csv
- **Has Image:** True/False from extractor tool
- **Has Video:** True/False from extractor tool
- **Network Size:** Network size at Post Timestamp DT from Connections.csv

5 Feature Engineering

To transform the dataset into a format required for the Random Forest and XGBoost classifier, feature engineering was implemented. This process involved decomposing raw post content into structural, semantic, and temporal variables while establishing a fair metric for performance across users of varying network sizes.

5.1 Post Content Features

The structure of a post serves as a primary signal for both the reader and the algorithm. The following features were engineered:

- **Post Content Length**: Depth of the post
- **Linebreak Count:** Readability and visual density
- **Emoji Count:** Emotional expression
- **Hashtag Count:** Classification and topic references
- **Hook Length:** Character count of the first line (up to the first newline character)
- **Link Count:** Number of external redirects inside the post content

5.2 Semantic and Sentiment Features

Beyond physical structure, the topical alignment and emotional resonance of a post was measured using advanced Natural Language Processing (NLP) models.

- **Semantic Alignment:** Utilizing the *paraphrase-multilingual-MiniLM-L12-v2* (Reimers, 2019) transformer, the cosine similarity between the post content and the author's profile summary was calculated. This feature serves as a simplified version for "Topical Authority," testing whether the algorithm rewards creators who remain within their established professional niche.
- **Sentiment Score:** Posts were analyzed using the *nlptown/bert-base-multilingual-uncased-sentiment* (Peirsman, 2020) model. By scoring content on a scale from negative (-1) to positive (+1), this feature captures the emotional tone, which is often a significant driver of viral engagement.

5.3 Post Timing

To test the hypothesis of the ideal posting time, **Hour of Day** (0..23) and **Day of Week** (0…6) were extracted from the standardized Post Timestamps.

5.4 Engagement and Performance (Target Variable)

A central challenge of this research was to compare performance across a cohort ranging from casual users to power users. To eliminate "Fame Bias," where large accounts naturally receive more engagement regardless of content quality, a weighted **Relative Engagement Score** (S_rel) was developed:

$$S\_rel = ((Likes * 1 + Comments * 3) / Network\,Size) * 1000$$

- **Weighting:** Comments are assigned a weight of 3, reflecting the higher algorithmic value and "Cost of Action" associated with starting a conversation compared to a simple like.
- **Normalization:** Dividing the weighted total by the **Network Size** creates a relative performance metric that focuses on how well a post converted a specific user's available audience.

5.5 Proxy Features for Algorithmic Signals

The literature review (Chapter 2) identified dwell time and comment quality as critical ranking signals in LinkedIn's algorithm. However, these metrics are not available in the LinkedIn Data Archive export. To address this limitation, the following features act as proxies:

5.5.1 Dwell Time Proxies

- **Hook Length:** The character count of the first line serves as a proxy for "stop the scroll" potential. A compelling hook increases the probability of sustained attention.
- **Post Content Length:** Longer posts require more reading time. If consumed fully, this signals content depth to the algorithm.
- **Linebreak Count:** Frequent paragraph breaks improve readability and reduce cognitive friction, potentially extending engagement duration.

5.6 Dataset distribution

An analysis of the feature engineered dataset was conducted to understand the statistical properties of the content, engagement, and creator-related features prior to modeling.

Post Content Length and Hook Length show strong right-skewed distributions, with most posts being short and a small number of very long posts forming a pronounced long tail. The median post length is 228 characters compared to a mean of 482, indicating the presence of extreme values.

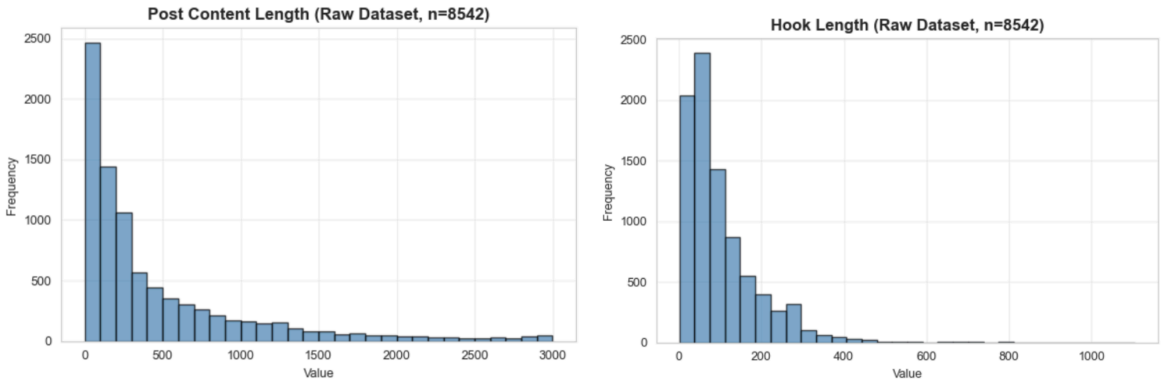

Figure 3 - Raw Dataset Distribution Post Content Length & Hook Length

Emoji Count, Hashtag Count, Linebreak Count and Link Count are heavily zero-inflated and highly skewed, with the majority of posts containing none of these elements, while a small subset uses them extensively. This suggests that such features represent deliberate stylistic choices rather than common posting behavior.

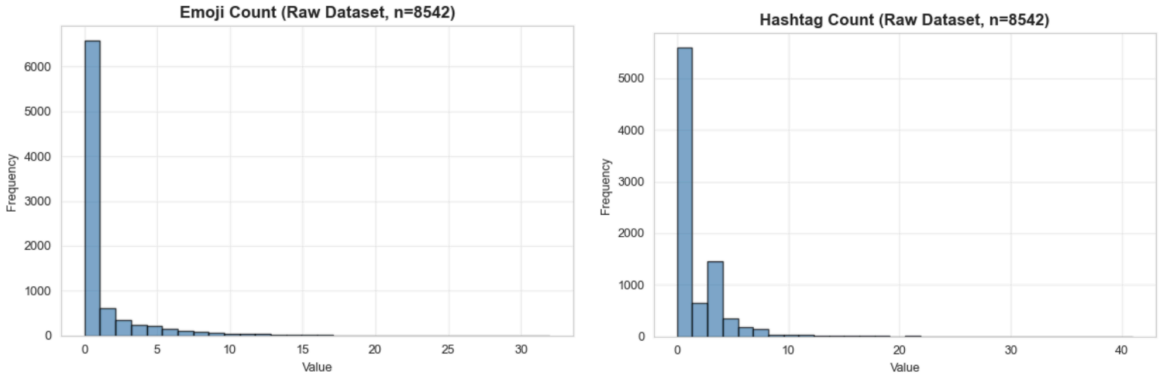

Figure 4 - Raw Dataset Distribution Emoji Count & Hashtag Count

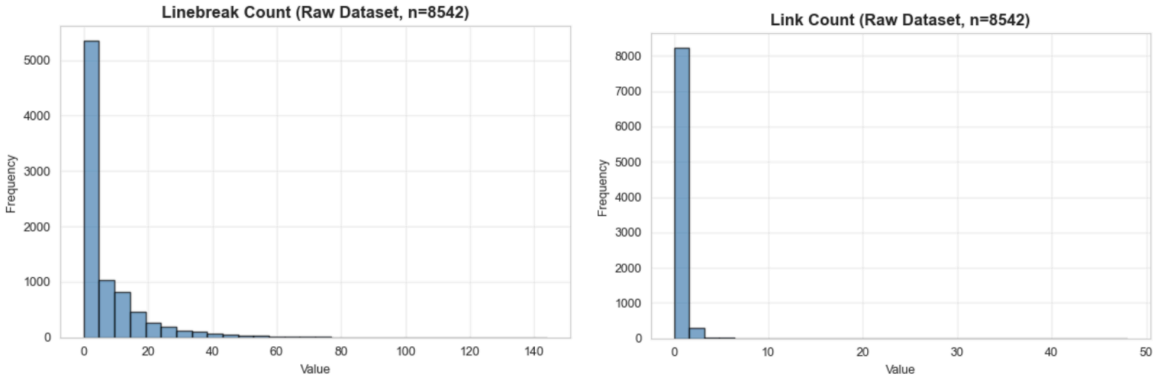

Figure 5 - Raw Dataset Distribution Linebreak Count & Link Count

Hour of Day and Day of Week are fairly distributed. Users more often posted in the morning hours and during work days. The window with the most published posts is Tuesday morning.

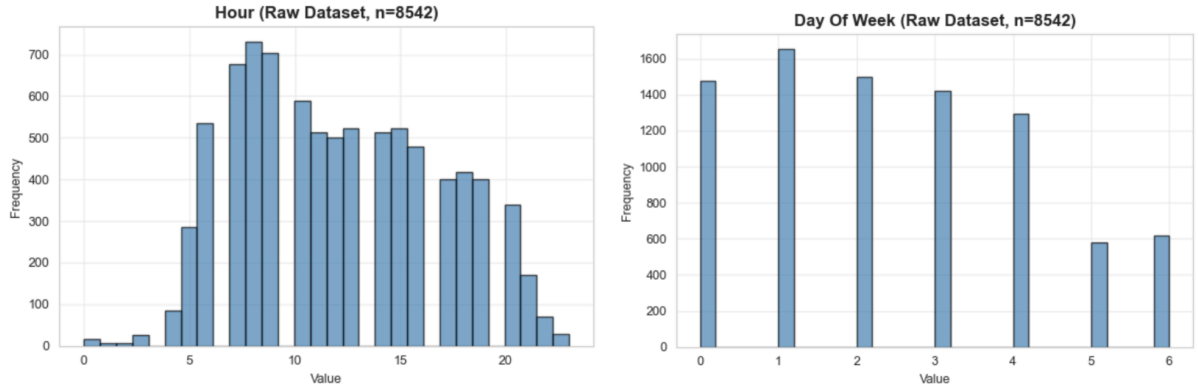

Figure 6 - Raw Dataset Distribution Hour & Day of Week

Network Size follows a right-skewed distribution, with most users maintaining moderate-sized networks and a smaller group possessing significantly larger audiences. Engagement metrics exhibit extreme skewness, with very low median values and a few outliers dominating the upper range. This heavy-tailed behavior is characteristic of social media data and motivated the use of outlier capping and relative engagement normalization.

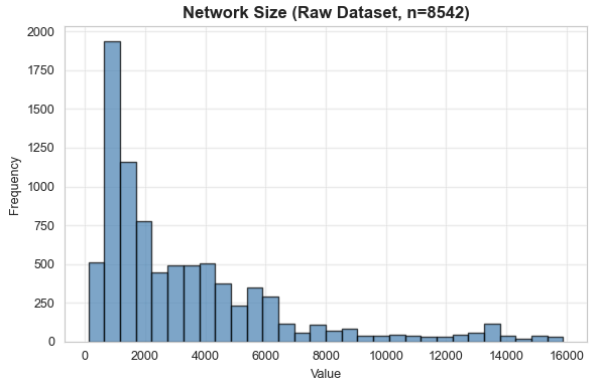

Figure 7 - Raw Dataset Distribution Network Size

Semantic Alignment displays a nearly symmetric distribution centered around moderate values, indicating consistent but variable topical alignment between posts and user profiles. Sentiment scores are slightly skewed toward positive values, suggesting a general tendency toward neutral-to-positive tone.

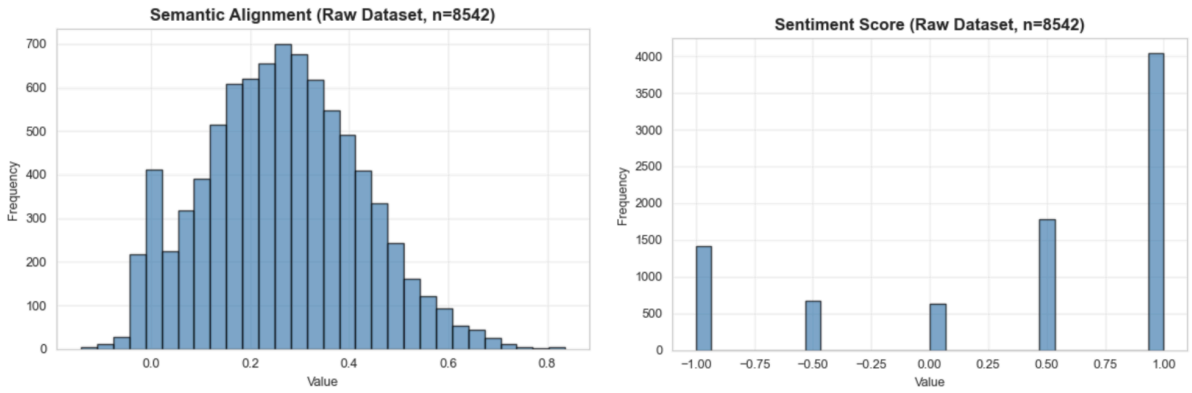

Figure 8 - Raw Dataset Distribution Semantic Alignment & Sentiment Score

The media features show a strong imbalance. While image posts are relatively common and provide meaningful variation, video posts are rare, limiting their predictive usefulness.

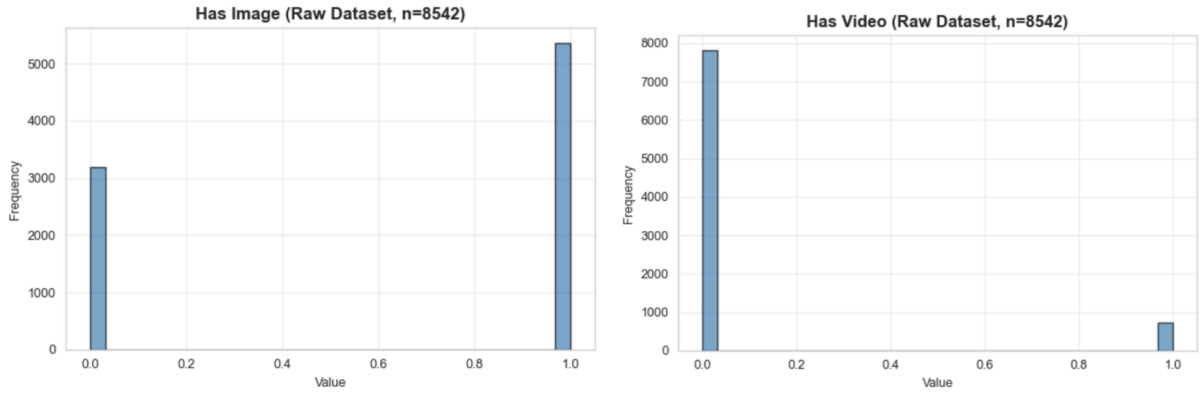

Figure 9 - Raw Dataset Distribution Has Image & Has Video

Overall, the data shows typical social media patterns, such as highly skewed engagement, many zero or low structural features, and varied posting behavior. These characteristics shaped the preprocessing, balancing, and normalization steps used in the analysis.

6 Training Dataset Generation

The transition from a raw, feature-rich dataframe to a training set requires careful decisions to ensure the model learns generalizable patterns rather than individual user identities. This chapter details the sampling and normalization strategies used to mitigate "Identity Bias" and handle the volatility in social media data.

6.1 Identity Bias and the Necessity of Sampling

In social media datasets, power users with thousands of historical posts can dominate the training process. If left unmanaged, a machine learning model may inadvertently learn to associate specific data points with high performance simply because that specific user is influential and dominant. To combat this, a three-step sampling strategy was implemented.

6.2 The Capping and Balancing Strategy

To ensure every participant contributed more equal to the model's training, the following constraints were applied during training dataset generation:

- **User Capping:** A maximum limit of **350 posts per user** was enforced. This ensures that the global model is not skewed by the high frequency of power users, forcing the algorithm to find structural commonalities across the entire 42-person cohort rather than over-indexing on a few influential authors.
- **Outlier Management (Capping):** Viral anomalies (posts that achieve engagement far beyond the statistical norm) can introduce significant noise. The **Relative Engagement Score** was capped at **150** (representing approximately the 99th percentile). This "Winsorization" technique prevents the model from being distracted by extreme outliers that are often driven by external factors or unpredictable viral trends.
- **Minimum Content Length:** Posts were required to have a minimal length of **50 characters**. This constraint was enforced to focus the model exclusively on in-depth, original posts, filtering out short, low-effort or reposted content.

The specific capping of 350 posts per user, Relative Engagement Score at 150 and minimum 50 character posts was not selected arbitrarily but was the result of an iterative methodological optimization process.

6.3 Final Training Set Characteristics

The final training set emphasizes the structural and semantic craftsmanship of the content. This ensures that the XGBoost model identifies features that are statistically significant across the entire spectrum of LinkedIn creators. The final dataset for training is 5124 posts large.

Following final features were created for model training to predict “Relative Engagement Score”:

- Post Content Length
- Hook Length
- Network Size
- Semantic Alignment
- Sentiment Score
- Emoji Count
- Hashtag Count
- Linebreak Count
- Link Count
- Hour of Day
- Day Of Week
- Has Image
- Has Video

7 Model Selection: Regression vs. Classification

An important decision in this project was determining how to mathematically represent "success." This chapter outlines the transition from continuous prediction to binary classification and the reasoning behind that shift.

7.1 The Limitations of Regression in Social Media Modeling

Initial experiments utilized a regression approach using Random Forest. The goal was to predict the exact Relative Engagement Score based on the model features. However, these models yielded poor results, with an R^2 value of approximately 0.20 - 0.30.

This highlighted the inherent nature of social media data:

- **Extreme Volatility:** Engagement is not a linear function; it often follows a power-law distribution where most posts receive modest attention while a tiny fraction achieves viral reach due to unpredictable external factors.
- **High Variance:** Two posts with identical structural features can have vastly different outcomes based on the specific composition of the feed at the moment of posting, a variable that cannot be captured through historical archive data.

7.2 Pivot to Binary Classification

The project shifted from predicting exact numbers to predicting relative performance via binary classification. Instead of asking "How many likes and comments will this get?", the model was to answer "Will this post perform better/worse than this specific user's typical content?"

This pivot transformed the target into a binary variable: **Is High Performing (1 or 0)**

7.3 Personalized Threshold to Map “Is High Performing”

To make the classification meaningful, a personalized median threshold was developed. A post is categorized as "High Performing" (1) if its engagement score exceeds the individual author's median score across their historical dataset. The median was selected for these methodological reasons:

- **Class Balance:** XGBoost performs optimally on balanced datasets. The median naturally creates a 50:50 split between high and low-performing posts, preventing the model from defaulting to majority-class predictions.
- **Outlier Robustness:** The median resists distortion from viral outliers, creating a stable baseline that rewards consistent quality over rare viral events."
- **Contextual Fairness:** It accounts for the "Baseline" of each user. For a casual user, 10 likes might be a significant success, whereas for a power user, it would be a failure. The median-based threshold treats both scenarios with equal mathematical weight.
- **Mitigating Identity Bias:** By defining success relative to the user's own history, the model is forced to learn what makes a post "better" than average, rather than simply learning that "Famous User A always gets high engagement."

7.4 Feature Correlation Analysis

Before proceeding to model training, a Pearson correlation analysis was conducted to quantify the linear relationships between the engineered features and the target variable, Is High Performing. This analysis serves as a statistical "pre-flight check," validating that the structural and media-based features possess predictive power before they are processed by the XGBoost algorithm.

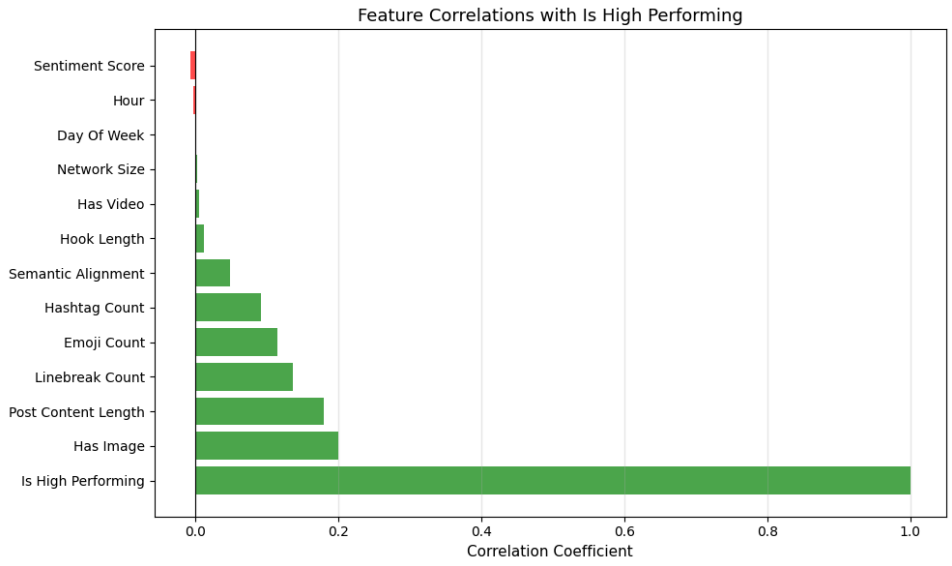

Figure 10 - Feature Correlations with Is High Performing

The correlations provide immediate insights into the "mechanics" of high-performing content:

- **The Visual Advantage:** Has Image (r = 0.2) emerged as the strongest linear correlate for high performance. Posts containing an image have a statistically significant higher probability of crossing the median engagement threshold.
- **The Content Length Signal:** Post Content Length (r = 0.18) and Linebreak Count (r = 0.14) follow closely. The algorithm identifies longer, well-formatted content as higher quality. The correlation with Emoji Count (r = 0.12) further supports the importance of visual density and readability.
- **The "Clean" Baseline:** Notably, Network Size shows a near-zero correlation with the binary target. This is a critical finding, confirming that the Relative Engagement Score successfully normalized performance and eliminated the linear effect of a user's follower count. The model can thus focus on *what* was posted, rather than *who* posted it.
- **Semantic Alignment's Weak Linear Signal:** Semantic Alignment (r = 0.05) shows a weak positive linear correlation. This suggests that while topical relevance is a factor, its true predictive power is not captured by a simple linear model and will require the non-linear capabilities of the XGBoost classifier for a full evaluation.

8 Model Training

Two primary learning models were evaluated to determine the optimal architecture for predicting post performance, with Random Forest serving as a baseline for comparison against XGBoost.

8.1 Random Forest vs. XGBoost

Random Forest was tested as a baseline method. Through hyperparameter tuning via grid search with 5-fold cross-validation, optimal parameters were identified favoring shallow trees and conservative splitting constraints. The model demonstrated stable training behavior with well-controlled overfitting, as confirmed by the learning curves showing consistent training accuracy while validation accuracy gradually improved with increased data size.

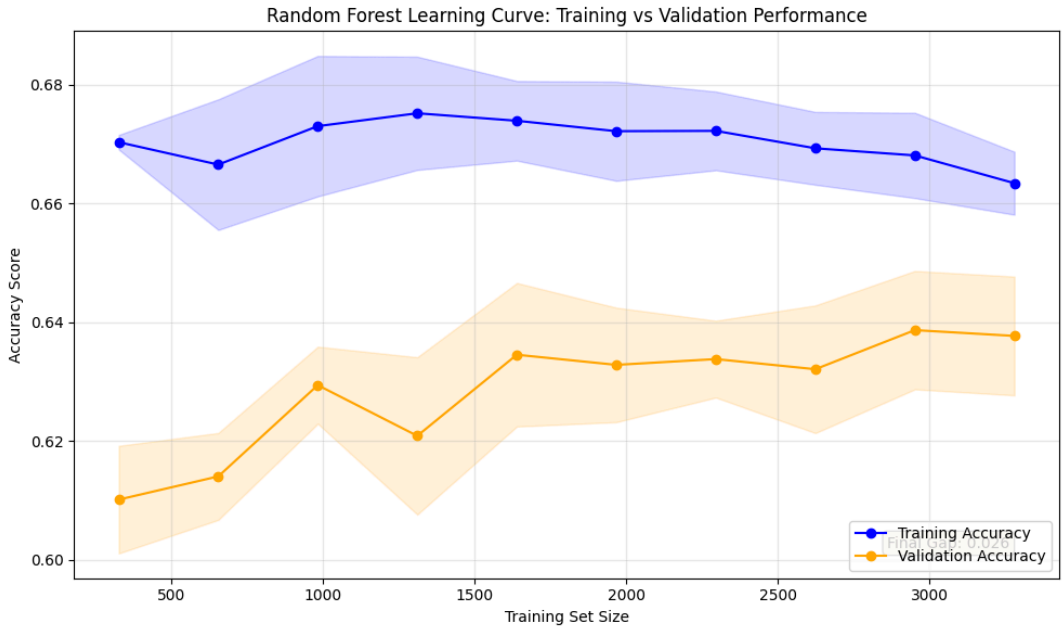

Figure 11 - Random Forest Learning Curve

XGBoost was evaluated as an alternative approach, utilizing gradient-based optimization with built-in regularization mechanisms. Hyperparameter tuning via grid search with 5-fold cross-validation yielded optimal parameters emphasizing strong regularization, aggressive subsampling, and a conservative learning rate with more estimators.

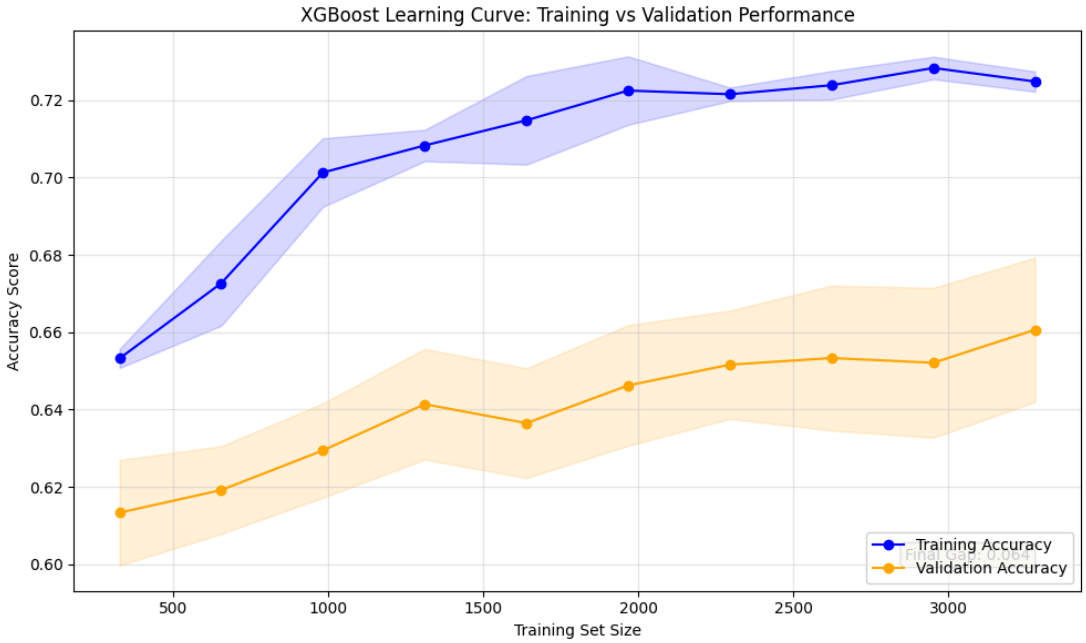

Figure 12 - XGBoost Learning Curve

While XGBoost exhibited a larger overfitting gap than Random Forest, the learning curves revealed that both training and validation accuracy continued to improve with additional training data, suggesting the model effectively leverages larger datasets.

Hyperparameter	**Random Forest**	**XGBoost**
Number of Estimators	50	700
Max Depth	5	4
Learning Rate	—	0.03
Min Samples Split / Min Child Weight	30	10
Min Samples Leaf	30	—
Max Features / Colsample by Tree	sqrt	0.7
Subsample	—	0.5
L1 Regularization (reg_alpha)	—	1.0
L2 Regularization (reg_lambda)	—	20
Gamma	—	1.0

Table 1 - Hyperparameter Comparison

Metric	**Random Forest**	**XGBoost**	**Difference**
Training Accuracy	0.6672	0.7230	+5.58%
Validation Accuracy	0.6406	0.6577	+1.71%
Overfitting Gap	0.0265	0.0653	+3.88%

Table 2 - Cross-Validation Performance

Metric	**Random Forest**	**XGBoost**	**Difference**
Accuracy	0.6585	0.6693	+1.08%
ROC-AUC	0.7193	0.7359	+1.66%

F1-Score	0.6622	0.6744	+1.22%
Precision	0.66	0.67	+1.00%
Recall	0.66	0.67	+1.00%
Val-Test Gap	-0.0179	-0.0115	+35.80%

Table 3 - Test Set Performance Comparison

Model Selection Decision: As shown in Table 3, XGBoost outperformed Random Forest across all evaluation metrics despite Random Forest exhibiting better overfitting control (Table 2). The learning curve analysis further indicated that XGBoost's performance had not yet plateaued, suggesting potential for additional gains with more training data. Given these consistent advantages in predictive performance, all subsequent analysis focused on XGBoost as the primary model.

8.2 XGBoost Hyperparameter Tuning

The final optimized XGBoost model hyperparameters, determined via RandomizedSearchCV with 5-fold cross-validation, prioritized strong regularization and enhanced generalization for the diverse user cohort.

The strategy involved:

- **Controlled Complexity:** The tree depth was constrained to max_depth: 4.
- **Conservative Learning:** A low learning_rate: 0.03 with n_estimators: 700 ensured slow, steady convergence, mitigating rapid overfitting.
- **Strong Regularization:** High L1/L2 penalties (reg_alpha: 1.0, reg_lambda: 20) and aggressive data subsampling (subsample: 0.5) were used.
- **Additional Stochasticity:** Feature sampling parameters (colsample_bytree: 0.7, colsample_bylevel: 0.5) and min_child_weight: 10 added further constraints to prevent fitting to specific feature interactions.

9 XGBoost Model Evaluation

The primary objective of the XGBoost model evaluation was to confirm that the model could reliably distinguish between high- and low-performing LinkedIn posts while maintaining a balanced accuracy across both classes.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.65      0.66       513
           1       0.66      0.69      0.67       512

    accuracy                           0.67      1025
   macro avg       0.67      0.67      0.67      1025
weighted avg       0.67      0.67      0.67      1025
```

9.1 Confusion Matrix

The model demonstrates slightly better performance at identifying high performing content (recall of 0.69) compared to low performing content (recall of 0.65). This asymmetry suggests the feature set captures success patterns more reliably than failure patterns.

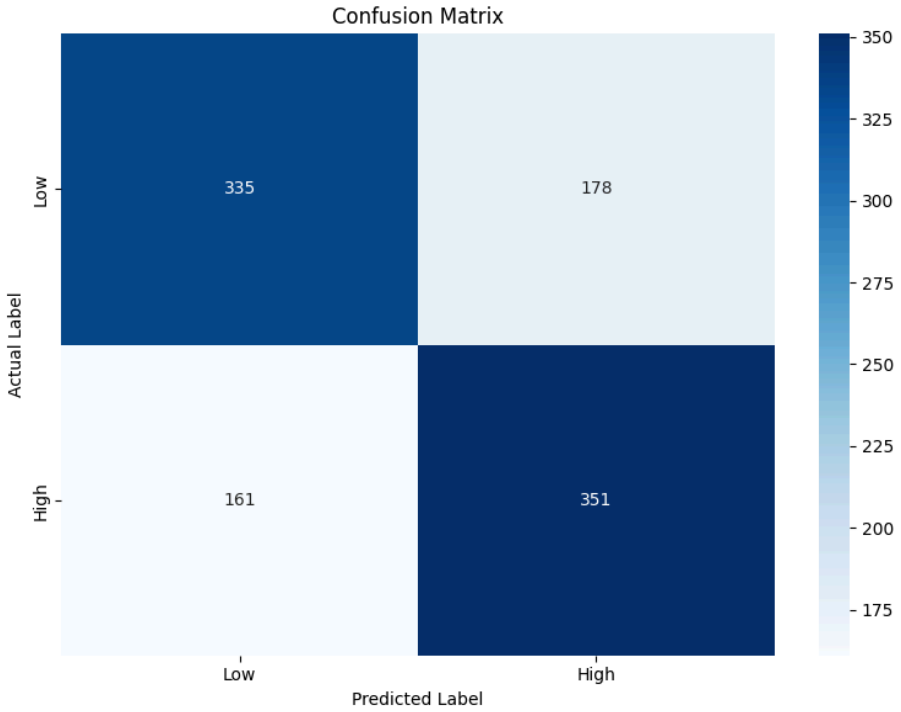

Figure 13 - Confusion Matrix

The relatively balanced error distribution (178 false positives vs. 161 false negatives) indicates the model is not systematically biased toward either class, despite the inherent unpredictability of social media engagement. The 686 correct predictions out of 1025 total cases yield the reported test accuracy of 0.67, confirming the model captures meaningful engagement signals while acknowledging the fundamental noise ceiling of this prediction task.

9.2 Feature Importance

The XGBoost feature importance analysis shows which features matter most overall. There is a clear dominance of visual and content-level features over timing and semantic variables.

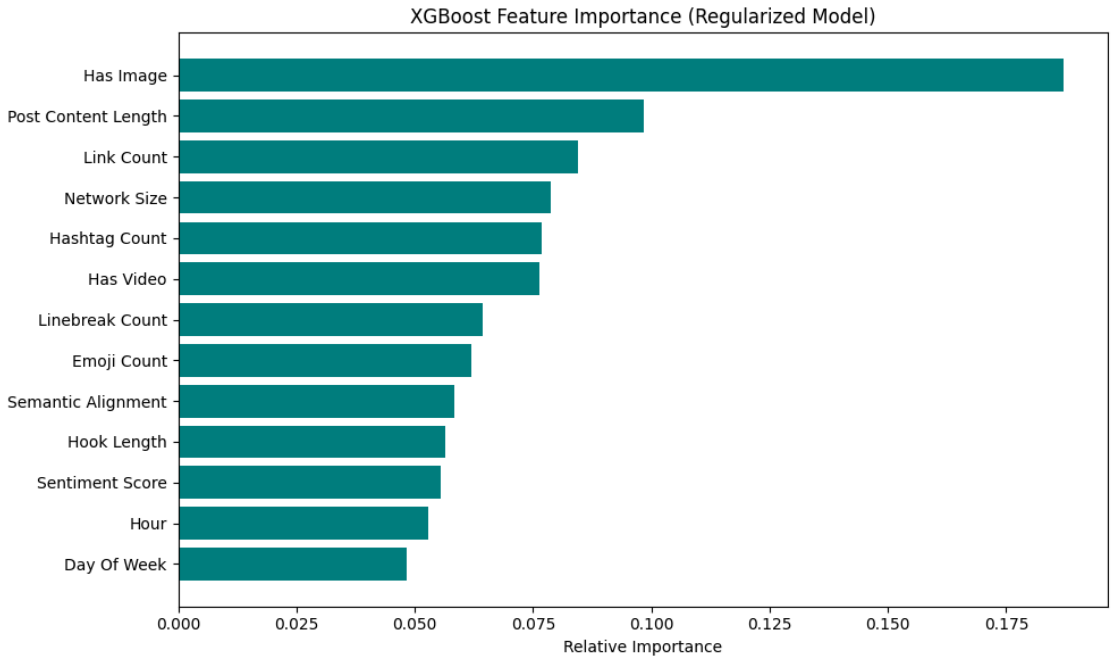

Figure 14 - Feature Importance

9.2.1 Visual Content Dominates

Has Image emerges as the single most important feature by a substantial margin (relative importance: ~0.19), nearly double the weight of the second-ranked feature. This indicates that visual presence functions as a primary gatekeeping signal in the LinkedIn algorithm, likely serving as a fast heuristic for content quality and user attention potential.

Has Video ranks sixth (~0.075), confirming that video content carries significant algorithmic weight, though substantially less than static images. The gap between image and video may also be a result from video being underrepresented in the data set (see Figure 9).

9.2.2 Content Substance Remains Critical

Post Content Length ranks second (~0.10), confirming that substantive, long-form content continues as a reliable quality signal. The model recognizes length as a proxy for depth, effort, and potential dwell time, making it a foundational engagement predictor alongside visual elements.

9.2.3 The Link Penalty and Structural Signals

Link Count ranks third (~0.085), indicating external links carry a meaningful algorithmic penalty. The presence of links likely signals content designed to redirect users off-platform, which LinkedIn's feed algorithm actively suppresses to maintain session duration.

Hashtag Count and **Network Size** cluster closely (0.08-0.082), representing important secondary signals. **Linebreak Count** and **Emoji Count** follow in the middle tier (~0.065-0.07), contributing to engagement prediction but not dominating outcomes.

9.2.4 Semantic and Emotional Signals

Semantic Alignment, **Hook Length**, and **Sentiment Score** occupy the third tier (0.055-0.06), indicating these features play supporting roles rather than primary drivers. The relatively low importance of Semantic Alignment contradicts the hypothesis that LinkedIn operates primarily through topic-matching retrieval, suggesting engagement potential and content quality outweigh pure semantic relevance.

9.2.5 Timing Remains Secondary

Hour of Day and **Day of Week** rank lowest (~0.05), confirming that posting time carries minimal weight in the engagement prediction function. This finding contradicts traditional social media optimization advice that emphasizes temporal targeting, suggesting LinkedIn's algorithm prioritizes content characteristics over scheduling tactics.

9.3 SHAP analysis

The SHAP (SHapley Additive exPlanations) analysis explains how and how much each feature contributes to predictions. The Global Importance (what matters most on average) and the Directional Distribution (how specific feature values change the outcome).

9.3.1 Global Feature Importance

The bar chart below ranks features by their average absolute impact on the model's predictions. Has Image and Post Content Length emerged as the dominant predictors, exhibiting substantially higher impact on model output than all other features. Network Size ranked third, highlighting the role of audience reach in determining post performance. A middle tier of

features—including Link Count, Hashtag Count, Hook Length, Sentiment Score, Semantic Alignment, and Linebreak Count—demonstrated moderate and relatively similar influence. Temporal features (Hour of Day, Day of Week) and formatting choices (Has Video, Emoji Count) contributed minimally to predictions, suggesting that content characteristics outweigh posting timing and multimedia choices in determining engagement outcomes.

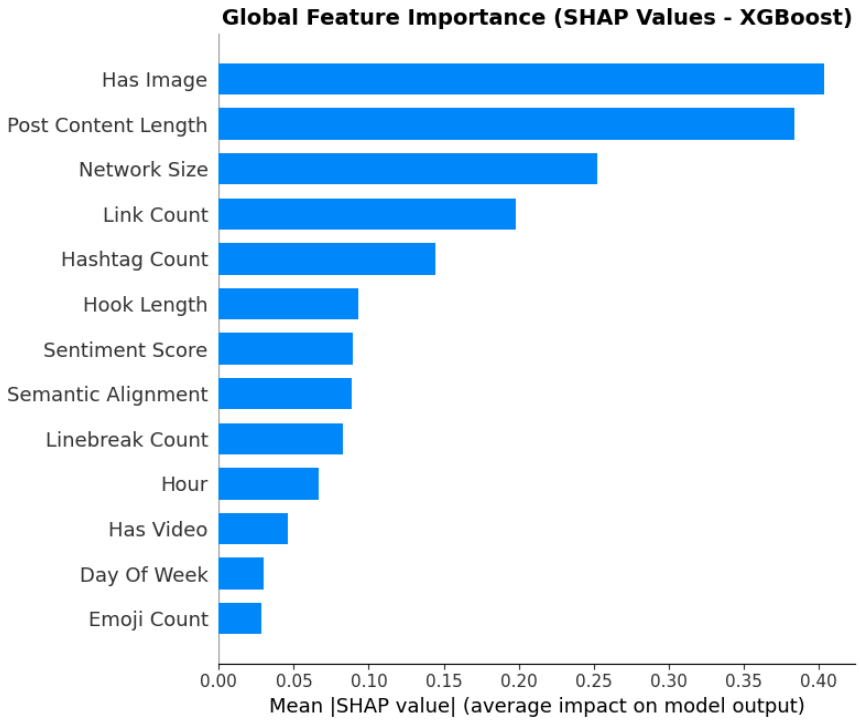

Figure 15 - SHAP Global Feature Importance

9.3.2 Directional Distribution (SHAP Beeswarm Plot)

The beeswarm plot illustrates how the *value* of a feature (high vs. low) pushes the prediction toward success or failure.

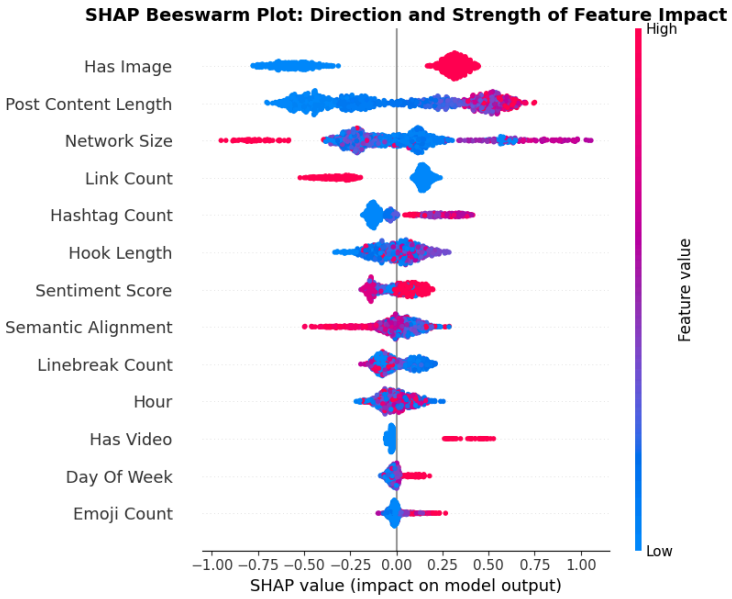

Figure 16 - SHAP Beeswarm Plot

9.3.3 Individual Feature Importance

Has Image is the most influential variable. Its dominant position indicates that the presence of an image is the primary influencer for engagement.

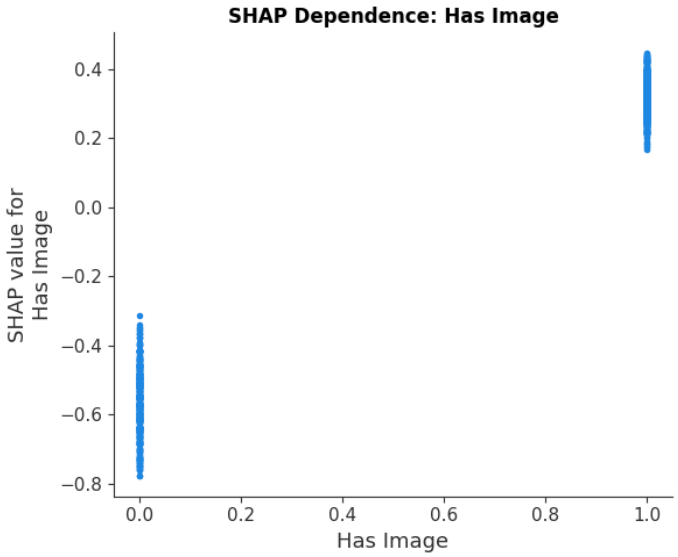

Figure 17 - SHAP Has Image

Post Content Length ranks second. This implies that engagement outcomes are materially influenced by content depth. The model identifies an "optimal range," starting from 500 characters length.

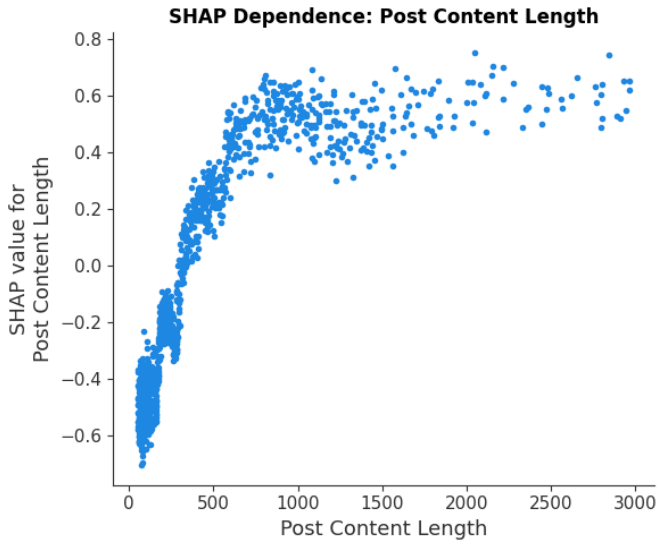

Figure 18 - SHAP Post Content Length

Network Size While author reach remains significant, its lower ranking relative to content features confirms that good content can effectively compete with, and often override, pure follower count.

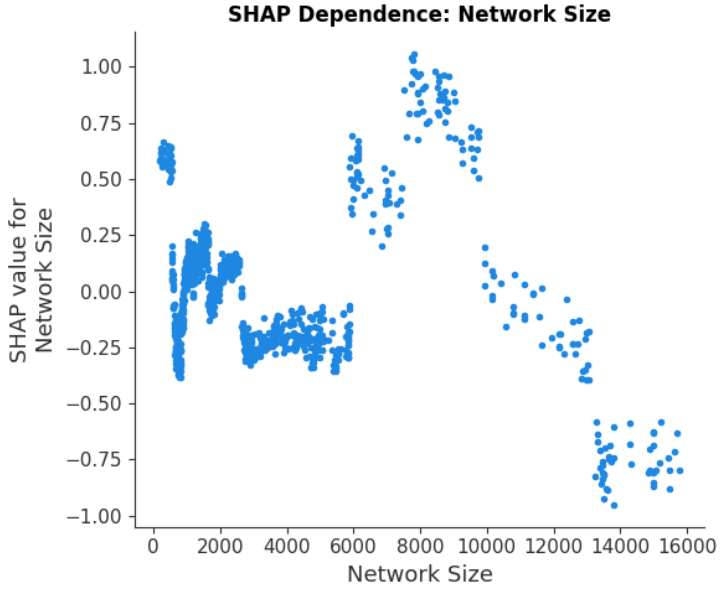

Figure 19 - SHAP Network Size

Link Count displays a clear negative skew. High values (more links) are associated with significantly lower SHAP values (reaching a minimum of **-0.65**). This provides empirical evidence that the platform actively suppresses posts that attempt to redirect users externally.

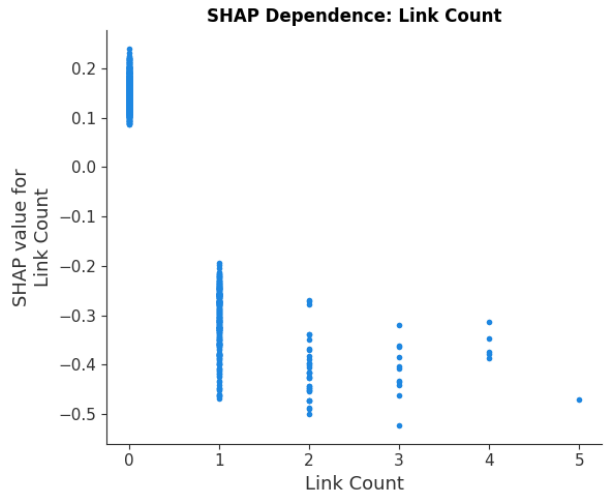

Figure 20 - SHAP Link Count

Hook Length shows high variance in its distribution. While it has a small mean effect, its range from -0.23 to +0.19 suggests it is a "precision lever." A well-optimized hook provides a strong positive push, while an over-long or missing hook acts as a definitive anchor on performance.

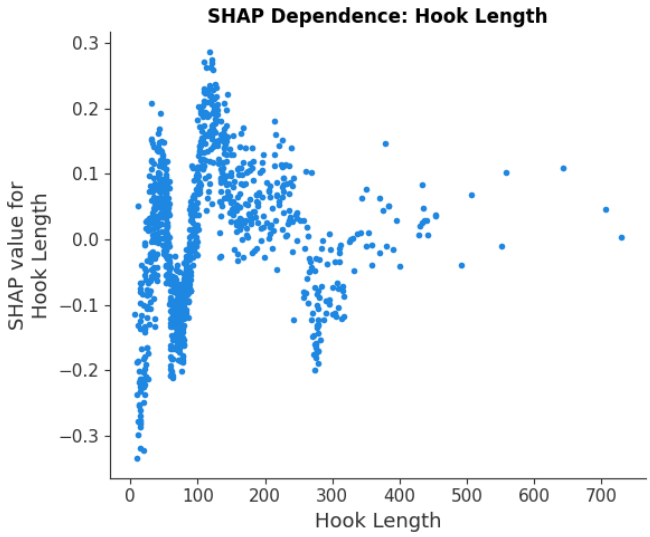

Figure 21 - SHAP Hook Length

Semantic Alignment has a subtle but structured positive impact. Higher alignment scores (relevance between the post and the author's professional niche) pushes towards a negative score before coming back positive again. The negative trend contradicts the original hypothesis.

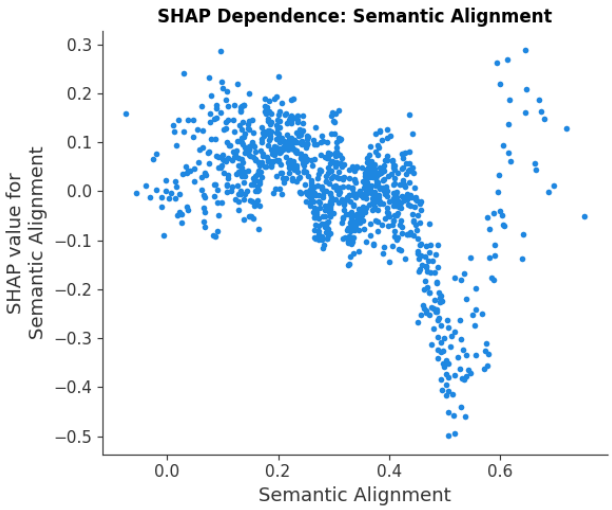

Figure 22 - SHAP Semantic Alignment

Both **Hour of Day** and **Day of Week** cluster tightly around zero. This is a significant finding for this thesis: it suggests that in an algorithmic feed environment, when you post is largely irrelevant compared to what you post.

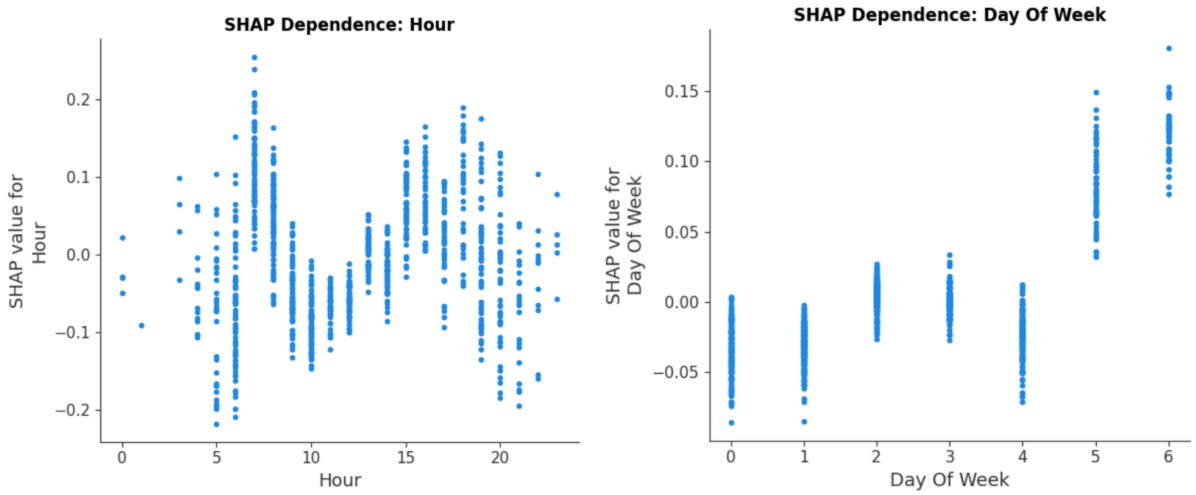

Figure 23 - SHAP Hour & Day

9.3.4 Research Synthesis

The model prioritizes content format and structure over traditional heuristics like posting time.

Driver Category	Impact Level	Primary Features
Primary Drivers	**High**	Has Image, Post Content Length, Network Size, Link Count
Secondary Drivers	**Medium**	Hashtag Count, Linebreak Count, Semantic Alignment, Sentiment Score
Tertiary Drivers	**Low**	Hour, Day, Emoji Count, Has Video

Table 4 - Drivers for Post Success

9.3.5 Implications for Content Strategy

The feature hierarchy found in the SHAP analysis reveals a clear content recommendation:

1. **Visual content is important** for algorithmic visibility
2. **Content length signals quality** and remains essential
3. **Avoid external links** to prevent algorithmic suppression
4. **Timing** optimization **is overrated** compared to content fundamentals
5. **Semantic matching nor sentiment is a primary ranking factor** despite speculation about topic-based feeds

10 Concept Drift: Old vs. New Algorithm

10.1 Background and Dataset Splitting

The final stage of this project involved a temporal split of the dataset to explore potential Concept Drift in LinkedIn's algorithm. By analyzing feature importance before and after March 2025, a period associated with reported algorithm updates. This analysis examines whether the platform's engagement rules may have evolved.

The dataset was split as follows:

- **Pre-March 2025** (until 2025-03-31): 4125 posts (80.5%)
- **Post-March 2025** (from 2025-04-01): 999 posts (19.5%)

10.2 Model Performance Across Time Segments

The models trained on each time segment showed notable performance differences:

Metric	All Data (for reference)	Pre-March 2025	Post-March 2025
Test Accuracy	0.669	0.689	0.615
ROC-AUC	0.736	0.751	0.657
F1-Score	0.674	0.694	0.617
Training Samples	4,099	3,300	799

Table 5 - Performance comparison across time segments

The Pre-March model achieved the strongest performance (68.9% accuracy, 0.751 ROC-AUC), while the Post-March model showed reduced predictive power (61.5% accuracy, 0.657 ROC-AUC). This performance gap may reflect:

1. **Sample size limitations**: The Post-March dataset contains only 999 posts (less than one-quarter of the Pre-March data), which increases variance in feature importance estimates and reduces the model's ability to learn robust patterns.
2. **Potential algorithmic changes**: If LinkedIn's algorithm has indeed shifted toward signals not captured in the current feature set, this would explain the reduced model performance.
3. **Time-related effects**: Seasonal effects, changes in user behavior, or shifts in the participant cohort's posting patterns could contribute to the observed differences.

10.3 SHAP-Based Evidence of Concept Drift

Despite the methodological limitations, the SHAP analysis reveals suggestive patterns in feature importance shifts:

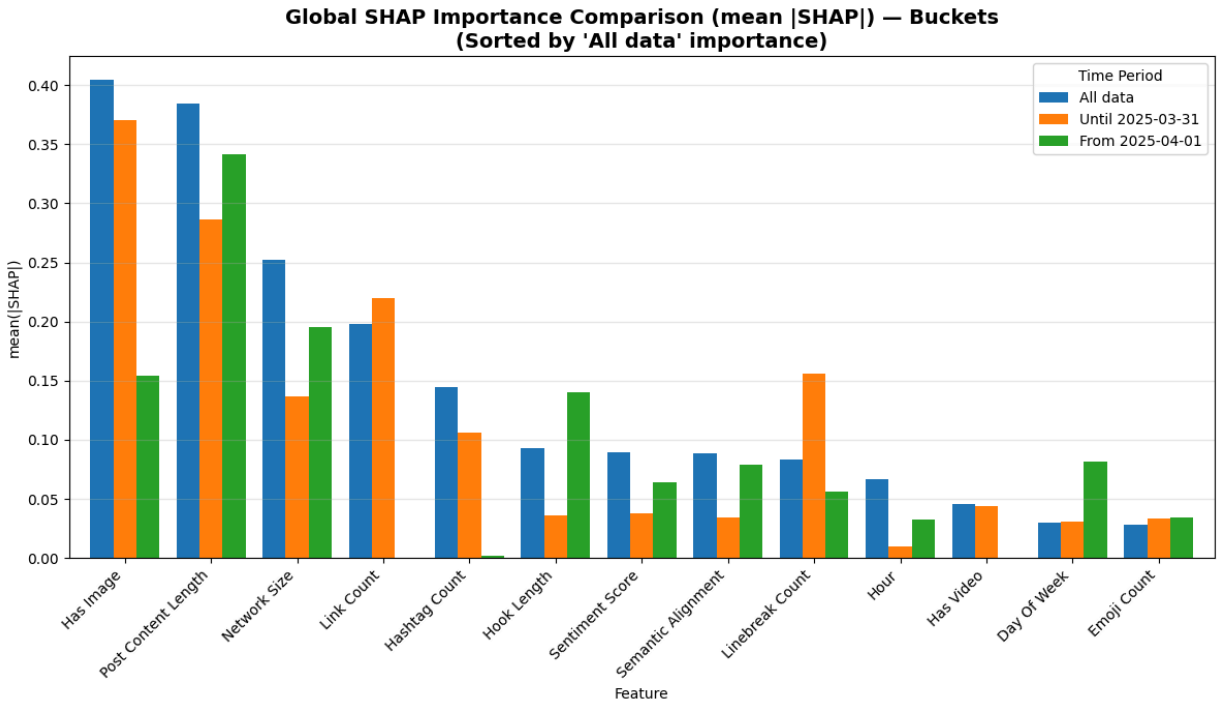

Figure 24 - Global SHAP Comparison All vs. Before vs. After

Feature	All Data (for reference)	Pre-March 2025	Post-March 2025	Δ (Absolute)	Δ (%)	Trend
Has Image	0.404	0.370	0.155	-0.216	-58.2%	↓↓↓
Post Content Length	0.384	0.286	0.341	+0.055	+19.2%	↑
Network Size	0.253	0.136	0.195	+0.059	+43.0%	↑↑
Link Count*	0.198	0.220	0.000	-0.220	-100.0%	↓↓↓
Hashtag Count*	0.144	0.106	0.002	-0.104	-98.2%	↓↓↓
Hook Length	0.093	0.036	0.140	+0.104	+284.2%	↑↑↑
Sentiment Score	0.089	0.038	0.064	+0.026	+68.2%	↑↑
Semantic Alignment	0.089	0.034	0.079	+0.045	+131.1%	↑↑↑
Linebreak Count	0.083	0.156	0.056	-0.100	-64.3%	↓↓
Hour	0.067	0.010	0.033	+0.023	+237.7%	↑↑↑
Has Video*	0.046	0.044	0.000	-0.044	-100.0%	↓↓↓

Day Of Week	0.030	0.031	0.081	+0.050	+161.0%	↑↑↑
Emoji Count	0.028	0.034	0.034	+0.000	+1.0%	→

Table 6 - Global SHAP Feature Comparison

**Features showing 0.000 importance in the Post-March period should be interpreted with caution due to small sample size.*

10.3.1 Observations with higher confidence

The following patterns appear more robust, as they involve features with meaningful importance values in both time periods:

- **Post Content Length remained the dominant predictor** in the Post-March period (0.341), suggesting that substantive content continues to drive engagement regardless of algorithm changes.
- **Network Size increased in relative importance** (0.136 → 0.195), indicating that creator authority may play a larger role in the newer algorithm.
- **Hook Length showed notable growth** (0.036 → 0.140), potentially supporting the hypothesis that compelling opening lines have become more critical for algorithmic distribution.
- **Semantic Alignment more than doubled** (0.034 → 0.079), which aligns with industry reports about LinkedIn's shift toward retrieval-based content matching.

10.3.2 Observations Requiring Caution

Several features showed extreme changes that may reflect data limitations rather than true algorithmic shifts:

- **Has Image dropped substantially** (0.370 → 0.155), but remained the third most important feature. The size of this decline may be amplified by the smaller sample size.
- **Link Count, Hashtag Count, and Has Video dropped to near-zero** importance. While this could indicate algorithmic changes, it more likely reflects insufficient variation in the smaller Post-March dataset to establish reliable importance estimates.
- **Timing variables (Hour of Day, Day Of Week)** showed large percentage increases but from very low base values. Their absolute importance remains modest (0.033 and 0.081 respectively), making strong conclusions premature.

10.3.3 Tentative Synthesis

The analysis provides suggestive patterns suggesting LinkedIn's algorithm may have evolved in the following directions:

1. **Increased emphasis on creator authority**: The growth in Network Size importance aligns with industry reports that LinkedIn now more heavily weights established creators.
2. **Growing role of semantic relevance**: The increase in Semantic Alignment importance, while modest in absolute terms, supports the hypothesis that LinkedIn's retrieval-based architecture rewards topical consistency.
3. **Hook quality as an emerging signal**: The substantial increase in Hook Length importance suggests that capturing initial attention may be increasingly critical.

However, these interpretations should be considered hypotheses for future investigation rather than definitive conclusions, given the methodological limitations outlined below.

10.4 Methodological Limitations

Several factors limit the strength of conclusions that can be drawn from this temporal analysis:

1. **Unbalanced sample sizes**: The Post-March dataset (999 posts) is less than one-quarter the size of the Pre-March dataset (4,125 posts), resulting in higher variance in all estimates.
2. **Reduced model performance**: The Post-March model's lower accuracy (61.5% vs 68.9%) and ROC-AUC (0.657 vs 0.751) indicate that the feature set explains less variance in the newer period. This could mean either (a) the algorithm now relies on signals not captured in our features, or (b) the smaller sample size prevents robust pattern detection.
3. **Single temporal cut point**: Using March 2025 as the split point is based on industry reports but may not precisely align with actual algorithm deployment dates.

10.5 Implications and Future Research

Despite these limitations, the concept drift analysis provides valuable directional insights:

For content creators: The consistent importance of Post Content Length across both periods reinforces that substantive content remains foundational. The potential growth in Hook Length and Semantic Alignment importance suggests that compelling openings and topical consistency may be increasingly valuable.

For future research: A more robust concept drift analysis would require:

- Larger Post-March dataset
- Rolling window analysis to detect gradual versus sudden shifts
- Controlled comparison of identical content types across time periods

The observed patterns align with LinkedIn's publicly stated direction toward expertise-driven distribution, but confirmation requires continued data collection and more rigorous temporal modeling.

11 Summary of Findings

This project successfully developed a machine learning framework to predict LinkedIn post performance by analyzing approximately 10,000 posts from 42 users. The XGBoost binary classification model achieved a test accuracy of 67%, representing a 17% improvement over random classification (50%). This demonstrates that engagement outcomes are systematically influenced by measurable content and creator-level features despite the inherent unpredictability of social media dynamics.

11.1 Primary Engagement Drivers

The presence of images showed the strongest global importance (SHAP: 0.404), confirming LinkedIn's algorithmic preference for media-rich content. Post content length ranked second (SHAP: 0.384), validating that substantive, well-crafted posts consistently outperform superficial updates. These findings establish a clear hierarchy: visual appeal serves as the initial attention gateway, while content depth sustains engagement.

11.2 Structural and Creator Signals

Network size demonstrated meaningful but secondary importance (SHAP: 0.253), indicating that creator authority reinforces but does not override content quality. The relative engagement scoring methodology successfully neutralized "fame bias," allowing the model to learn content patterns rather than creator identities. Link count exhibited a consistent negative impact (SHAP: 0.198), empirically validating LinkedIn's "walled garden" strategy of suppressing posts with external redirects.

11.3 Semantic and Sentiment Features

Semantic alignment between user profiles and post content showed a positive but moderate effect (SHAP: 0.089). While topical consistency contributes to performance, it functions as a reinforcing signal rather than a primary driver. Sentiment similarly showed limited independent predictive power (SHAP: 0.089), suggesting that emotional tone matters less than structural craftsmanship and visual richness.

11.4 Posting Time Myth

Perhaps the most actionable finding contradicts conventional social media wisdom: posting time (hour of day and day of week) demonstrated minimal predictive power in the overall analysis (SHAP: 0.067 and 0.030 respectively). This suggests that LinkedIn's algorithmic feed prioritizes content quality over temporal optimization, fundamentally challenging the "ideal posting time" advice prevalent in marketing literature.

11.5 Concept Drift

A temporal split analysis (Pre-March 2025 vs. Post-March 2025) was conducted to explore potential algorithmic evolution. While the analysis revealed suggestive patterns, the findings should be interpreted with caution due to methodological limitations:

Observations with higher confidence:

- Post Content Length remained a dominant predictor across both periods, confirming that substantive content is consistently rewarded.
- Network Size and Semantic Alignment showed increased importance in the Post-March period, tentatively supporting industry reports about LinkedIn's shift toward creator authority and retrieval-based content matching.
- Hook Length appeared to gain importance, suggesting that compelling opening lines may be increasingly critical.

Limitations affecting interpretation:

- The Post-March dataset (999 posts) was substantially smaller than the Pre-March dataset (4,125 posts), increasing variance in feature importance estimates.
- The Post-March model achieved lower accuracy (61.5% vs. 68.9%) and ROC-AUC (0.657 vs. 0.751), indicating either algorithmic changes toward unmeasured signals or insufficient data for robust pattern detection.
- Several features (Link Count, Has Video) dropped to near-zero importance in the Post-March period, which likely reflects sample size limitations rather than true algorithmic changes.

11.6 Methodological Contributions

The project successfully mitigated identity bias through a 3-step approach:

1. Capping posts per user at 350
2. Maintaining 50:50 class balance within each creator's contribution
3. Normalizing engagement through the relative scoring methodology.

11.7 Model Generalization

The learning curve analysis confirmed appropriate regularization. The final training-validation gap of 6.53% and validation-test gap of -1.2% demonstrate that the model captures genuine engagement patterns rather than overfitting to training data. The 67% accuracy, while modest in absolute terms, represents meaningful predictive power in a domain characterized by high stochasticity, outperforming random classification by 17%.

12 Future Outlook

Expand Data Generalization: Conduct an expanded data collection effort beyond the author's professional network to include a more diverse and geographically representative cohort, thereby improving model generalization across different industries and professional segments.

Integrate Advanced Engagement Metrics: Secure access to advanced, proprietary engagement metrics (such as impressions, saves, shares, content of comments) through the LinkedIn API to allow the model to capture a richer signal quality and more accurately predict virality and algorithmic amplification.

Strengthen Concept Drift Analysis: Undertake a continuous data collection over an extended period (12–18 months post-algorithm change) to acquire a statistically robust Post-March dataset, enabling reliable statistical testing of observed concept drift and temporal shifts in feature importance.

Cold Start Problem: Implement GroupKFold cross-validation to test model generalization to completely unseen creators. This would validate whether the model learns universal engagement principles or remains dependent on having historical data for each user.

Real-Time Prediction: Deploy the model as a pre-posting advisory tool that provides creators with predicted performance scores and actionable optimization suggestions before publication.

References

brightdata. (-, - -). *brightdata*. LinkedIn Posts Dataset. Retrieved 1 29, 2026, from https://brightdata.com/products/datasets/linkedin/posts

Fox, T. (2023, December 6). *LinkedIn Post Date Extractor: How to See the Exact Date of a LinkedIn Post*. Trevor Fox. Retrieved January 29, 2026, from https://trevorfox.com/linkedin-post-date-extractor.html

Johnson, C. (2025, 11 18). *Grab A Cup of 360 Brew, LinkedIn's New Algorithm*. LinkedIn. Retrieved 1 29, 2026, from https://www.linkedin.com/pulse/grab-cup-360-brew-linkedins-new-algorithm-chad-johnson-rckjc/

Lämmler, R. (2025, 12 16). *LinkedIn*. Donate your data. Retrieved 1 29, 2026, from https://www.linkedin.com/posts/rlaemmler_i-need-your-help-im-back-in-school-activity-7406606552159375362-Q4rK

Malik, D. (2025, 03 11). *100 Essential LinkedIn Statistics and Facts for 2025: Your Guide to the Professional Network*. LinkedIn. Retrieved 1 29, 2026, from https://www.linkedin.com/pulse/100-essential-linkedin-statistics-facts-2025-your-guide-dilawar-malik-pog9f/

Newberry, C., & Christison, C. (n.d.). *How the LinkedIn algorithm works in 2025*. Hootsuite Blog. Retrieved January 29, 2026, from http://blog.hootsuite.com/linkedin-algorithm/

Peirsman, Y. (2020, 2 14). *nlptown/bert-base-multilingual-uncased-sentiment · Hugging Face*. Hugging Face. Retrieved January 29, 2026, from https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

Ramanujam, S. S., Alonso, A., Kataria, S., Dangi, S., Gupta, A., Tiwana, B. S., Somaiya, M., Simon, L., & Byrne, D. (2025, 10 16). *Large Scale Retrieval for the LinkedIn Feed using Causal Language Models*. Arxiv. Retrieved 1 29, 2026, from https://arxiv.org/html/2510.14223v1

Reimers, N. (2019, August 27). *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 · Hugging Face*. Hugging Face. Retrieved January 29, 2026, from https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

List of images

List of tables

Appendices

Appendix 1: XGBoost Classification Model with Enhanced Regularization & Diagnostics

```
# Load the dataset
df_ml = pd.read_csv(INPUT_FILE)

# Define Features (X) and Binary Target (y)
features = [
    'Network Size',
    'Post Content Length',
    'Hook Length',
    'Semantic Alignment',
    'Sentiment Score',
    'Emoji Count',
    'Hashtag Count',
    'Linebreak Count',
    'Link Count',
    'Hour',
    'Day Of Week',
    'Has Image',
    'Has Video'
]

X = df_ml[features]
y = df_ml['Is High Performing']

print(f"Total rows: {len(df_ml)}")

# Check class balance
print("\nClass Distribution:")
print(y.value_counts(normalize=True))
class_ratio = y.value_counts()[0] / y.value_counts()[1]
print(f"Class ratio (0:1): {class_ratio:.2f}")

# Train/Test/Split with stratification
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42, stratify=y)
```

```
# MODELING: XGBOOST WITH STRONG REGULARIZATION
print("\nTraining XGBoost with Enhanced Regularization...")

# Aggressive regularization to close the overfitting gap
param_grid = {
    # Even simpler trees
    "max_depth": [2, 3, 4],  # Focus on shallower trees
    "min_child_weight": [10, 15, 20, 25],  # Higher minimum samples

    # Slower learning
    "learning_rate": [0.01, 0.02, 0.03],
    "n_estimators": [200, 300, 500, 700],

    # Stronger regularization
    "reg_alpha": [1.0, 2.0, 5.0, 10.0],  # Stronger L1
    "reg_lambda": [5, 10, 15, 20],  # Stronger L2

    # More aggressive sampling
    "subsample": [0.5, 0.6, 0.7],
    "colsample_bytree": [0.5, 0.6, 0.7],
    "colsample_bylevel": [0.5, 0.6, 0.7],

    # Higher split cost
    "gamma": [1.0, 2.0, 3.0, 5.0],

    # Handle class imbalance
    "scale_pos_weight": [1, class_ratio, class_ratio * 1.5],
}

# Use StratifiedKFold
from sklearn.model_selection import StratifiedKFold
cv_strategy = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

search = RandomizedSearchCV(
    XGBClassifier(
        objective="binary:logistic",
        eval_metric="logloss",
        tree_method="hist",
        random_state=42,
        n_jobs=-1
    ),
```

```
    param_distributions=param_grid,
    n_iter=100,  # More iterations for better search
    cv=cv_strategy,
    scoring="accuracy",
    n_jobs=-1,
    verbose=1,
    random_state=42,
    return_train_score=True
)


search.fit(X_train, y_train)
best_model = search.best_estimator_

# Detailed performance analysis
cv_results = pd.DataFrame(search.cv_results_)
best_idx = search.best_index_
train_score = cv_results.loc[best_idx, 'mean_train_score']
val_score = cv_results.loc[best_idx, 'mean_test_score']
overfitting_gap = train_score - val_score

print(f"\n{'='*60}")
print(f"Best Hyperparameters:")
print(f"{'='*60}")
for param, value in search.best_params_.items():
    print(f"  {param}: {value}")
print(f"\nCross-Validation Performance:")
print(f"  Training Accuracy:   {train_score:.4f}")
print(f"  Validation Accuracy: {val_score:.4f}")
print(f"  Overfitting Gap:     {overfitting_gap:.4f}")
if overfitting_gap < 0.05:
    print(f"  ✓ Overfitting well controlled!")
elif overfitting_gap < 0.10:
    print(f"  ⚠ Moderate overfitting - consider more regularization")
else:
    print(f"  ✗ Significant overfitting detected")
print(f"{'='*60}")

# EVALUATION
y_pred = best_model.predict(X_test)
y_pred_proba = best_model.predict_proba(X_test)[:, 1]
```

```
from sklearn.metrics import roc_auc_score, f1_score, precision_recall_curve, auc

test_acc = accuracy_score(y_test, y_pred)
test_auc = roc_auc_score(y_test, y_pred_proba)
test_f1 = f1_score(y_test, y_pred)

print(f"\nTest Set Performance:")
print(f"  Accuracy:  {test_acc:.4f}")
print(f"  ROC-AUC:   {test_auc:.4f}")
print(f"  F1-Score:  {test_f1:.4f}")

# Check if test performance matches validation
val_test_gap = val_score - test_acc
print(f"  Val-Test Gap: {val_test_gap:.4f}")
if abs(val_test_gap) < 0.02:
   print(f"  ✓ Good generalization to test set!")

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

# VISUALIZATIONS

# 1. Feature Importance
importances = best_model.feature_importances_
indices = np.argsort(importances)
plt.figure(figsize=(10, 6))
plt.title('XGBoost Feature Importance (Regularized Model)')
plt.barh(range(len(indices)), importances[indices], color='teal', align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.tight_layout()
plt.show()

# 2. Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
           xticklabels=['Low', 'High'], yticklabels=['Low', 'High'])
plt.title('Confusion Matrix')
plt.ylabel('Actual Label')
plt.xlabel('Predicted Label')
plt.tight_layout()
```

```
plt.show()

# 3. Learning Curve
from sklearn.model_selection import learning_curve

print("\nGenerating Learning Curve...")
train_sizes, train_scores, val_scores = learning_curve(
   best_model,
   X_train,
   y_train,
   train_sizes=np.linspace(0.1, 1.0, 10),
   cv=5,
   scoring='accuracy',
   n_jobs=-1,
   random_state=42
)

train_mean = np.mean(train_scores, axis=1)
train_std = np.std(train_scores, axis=1)
val_mean = np.mean(val_scores, axis=1)
val_std = np.std(val_scores, axis=1)

plt.figure(figsize=(10, 6))
plt.plot(train_sizes, train_mean, label='Training Accuracy', color='blue', marker='o')
plt.fill_between(train_sizes, train_mean - train_std, train_mean + train_std,
alpha=0.15, color='blue')
plt.plot(train_sizes, val_mean, label='Validation Accuracy', color='orange',
marker='o')
plt.fill_between(train_sizes, val_mean - val_std, val_mean + val_std, alpha=0.15,
color='orange')

gap = train_mean[-1] - val_mean[-1]
plt.text(0.95, 0.05, f'Final Gap: {gap:.3f}',
        transform=plt.gca().transAxes,
        bbox=dict(boxstyle='round', facecolor='wheat', alpha=0.5),
        verticalalignment='bottom', horizontalalignment='right')

plt.xlabel('Training Set Size')
plt.ylabel('Accuracy Score')
plt.title('Learning Curve: Training vs Validation Performance')
plt.legend(loc='lower right')
plt.grid(True, alpha=0.3)
```

```
plt.tight_layout()
plt.show()
```

Appendix 2: Github Repo

https://github.com/retolaemmler/linkedin-post-success-predictor