Per-Tier Engagement Classification on LinkedIn

LinkedIn Engagement Prediction — TrendPilot

Notebook: `11f_per_tier_models.ipynb` **Date:** February 2026

1. Overview

This experiment builds four independent binary classifiers — one per creator-size tier — to predict whether a LinkedIn post will perform above or below average for its segment. Rather than applying a single global model and threshold to all creators, each tier's model learns from its own training subset and classifies posts relative to that tier's median engagement rate.

Three algorithms are evaluated per tier (Random Forest, XGBoost, LightGBM), and the best-performing model is selected independently for each tier. All 71 input features are content-only — no audience-size signals are passed to any model.

Core result: Every tier's best model achieves F1 > 0.500 (random baseline), confirming genuine content signal exists within each creator segment. The weighted average Macro F1 across all four tiers is **0.7199**, a lift of **+0.2199** over chance.

2. Problem Statement

2.1 The Limitation of a Global Threshold

Engagement rate normalises raw engagement for audience size:

```
engagement_rate = (reactions + comments) / (followers / 1,000)
```

This expresses total engagement as engagements per 1,000 followers. Because the denominator is `followers / 1,000` rather than raw `followers`, the metric can produce values far above 1 — there is no upper bound. For a creator with 500 followers, the denominator is 0.5, so 60 total engagements yields an ER of 120. For a creator with 2,000 followers, the same 60 engagements yields an ER of 30. These are not anomalous results; they reflect the actual engagement density of small, tight-knit communities. The micro tier median of 22.216 in this dataset is a direct consequence of these small denominators and closely-knit audiences. The formula is always well-defined because follower counts are always positive, but the practical range across creator sizes is enormous — from fractions of a unit for large broadcast accounts to tens or hundreds for small personal ones.

This structural range makes creators of different sizes nominally comparable in direction — more is better, less is worse — but not in absolute magnitude. Even after this normalisation, the per-follower rate is not equivalent across creator tiers. Micro-creators (fewer than 10,000 followers) tend to have small, niche, and highly engaged audiences — their engagement rates are structurally higher. Large accounts (more than 200,000 followers) reach broad but more passive audiences — their rates are structurally lower.

A single global median threshold applied to all creators encodes this structural difference into the class boundary. A large-account post is labelled "below average" simply because the global median sits far above what large accounts typically produce — not because the post underperformed relative to its own segment.

2.2 The Per-Tier Solution

The correct framing is: "*Is this post above average for this type of creator?*" Each creator segment has its own engagement norm, and the class boundary should reflect that norm.

This experiment implements this directly. For each of four follower-based tiers, a separate binary classifier is trained using a threshold derived from that tier's own training-subset median. A micro-creator post is evaluated against the micro-creator median. A large-account post is evaluated against the large-account median. The random baseline remains exactly 0.500 for every tier by construction.

2.3 Why the Tier Median, Not the Mean

The class threshold for each tier is the **median** engagement rate of that tier's training posts — not the mean (arithmetic average). This distinction matters.

Engagement rate distributions are strongly right-skewed. The majority of posts earn moderate engagement, but a small number go "viral" within their audience and produce values far above the typical range. A post with 500 reactions from a 2,000-follower account produces an ER of 250 — roughly ten times the micro median. If the mean were used as the class threshold, this single outlier would pull it upward substantially, making it harder for normal above-average posts to be labelled Class 1. The mean represents "average performance inflated by occasional exceptional posts," not "typical post performance."

The median is the 50th percentile — it is the point where exactly half the training posts fall above and half fall below, regardless of how extreme the outliers are. This gives the median three properties that make it the right choice here:

1. **Robustness.** Viral outliers do not distort it. The threshold represents genuinely typical performance, not a target inflated by rare events.
2. **Guaranteed balance.** By definition, the median always produces an approximately 50/50 class split in the training data. This means the random baseline is exactly 0.500 for every tier, making model lift directly comparable across segments.
3. **Stability.** Small changes in the composition of the training set (which posts happen to appear in it) do not dramatically shift the median threshold, whereas the mean can move substantially if a few outliers are added or removed.

Using the mean would also produce unequal class splits across tiers — tiers with more extreme outliers would have a higher proportion of below-average posts, complicating cross-tier comparison. The median eliminates this problem by construction.

2.4 Why This Framing Matters

For a content scoring system intended to help creators understand whether their content is strong relative to their peers, the per-tier approach is the only fair one. A 5,000-follower creator should not be penalised because their engagement rate is below the macro-influencer median. By confining comparisons within segments, each creator is measured against their actual competitive set.

3. Dataset

3.1 Overview

The dataset contains **772 LinkedIn posts** from 495 unique authors (mean 1.56 posts per author). Posts span a wide range of creator sizes, from accounts under 1,000 followers to accounts above 2,500,000 followers. The full dataset is split 80/20 (stratified) into training and test sets before any per-tier processing.

3.2 Follower Tier Distribution

A **tier** is a grouping of LinkedIn creators based on their follower count, using fixed breakpoints that correspond to standard influencer-marketing categories. Tier membership determines which model evaluates a post and which median threshold is used as the class boundary. The four tiers used in this experiment are:

Micro (< 10,000 followers): Individual professionals, subject-matter enthusiasts, early-career practitioners, and community builders. The creator-follower relationship at this scale is personal — followers often know the creator or discovered them through a specific niche. Posts feel like direct conversation. Comments tend to come from genuine peers rather than passive scrollers, which is why raw engagement counts can translate into high per-follower rates even from modest absolute numbers (e.g. 200 reactions from a 3,000-follower account yields ER = 66.7). This tier dominates the dataset (345 posts, 45% of total).

Small (10,000 – 50,000 followers): Emerging professional voices who have built an established niche audience. These creators have enough followers to be recognised within their field but still maintain enough audience proximity for personal storytelling to drive engagement. The audience is larger and somewhat more passive than micro, but the creator-follower relationship is still meaningfully personal. This is the most common profile of an "active LinkedIn creator" who produces original content regularly.

Medium (50,000 – 200,000 followers): Established professional voices — typically mid-career to senior professionals, consultants, or specialists with recognised industry authority. At this scale, posts are reaching audiences that include many unfamiliar followers, shifting the dynamic from community conversation toward broadcast communication. Content needs stronger structural hooks to arrest scrolling behaviour. The post is competing for attention from a more diverse and less captive audience.

Large (> 200,000 followers): Senior executives, prominent founders, well-known public figures, or major institutional accounts. At this scale the follower relationship is primarily one-directional. Audiences follow for professional authority, not personal connection. Engagement rates are structurally lower because the denominator (followers/1,000) is large. A large account with 500,000 followers needs 167 reactions just to reach ER = 0.337 — the large-tier median. A post that "goes viral" for a large account in absolute terms may still only narrowly clear the tier median in per-follower terms.

Tier	Follower Range	n_train	n_test
micro	< 10,000	268	77
small	10,000 – 50,000	184	41
medium	50,000 – 200,000	76	16
large	> 200,000	89	21

Micro and small creators together account for 74% of the dataset, consistent with the general LinkedIn creator population. Medium and large tiers have substantially smaller samples, which is the binding constraint on their model reliability.

3.3 Per-Tier Median Engagement Rates

The tier median is computed from the training subset only and serves as the binary class boundary for that tier's model:

Tier	Tier Median ER (engagements per 1k followers)
micro	22.216
small	3.340
medium	1.723
large	0.337

The 66× gap between the micro median (22.216) and the large median (0.337) quantifies just how structurally different these segments are. A micro-creator whose post earns 22 engagements per 1,000 followers is performing at their segment's median. A large account at the same rate is performing at roughly 66× their own median — an extraordinary outlier. This heterogeneity is precisely why a single global threshold produces unfair classifications.

4. Feature Set

4.1 What Is Included

Each model receives **71 content features** covering the following categories:

Category	Example Features
Text quality	`text_lexical_diversity`, `text_difficult_words_count`, `text_avg_sentence_length`
Readability	`readability_flesch_kincaid`, `readability_gunning_fog`
Sentiment	`sentiment_compound`, `sentiment_x_readability`
Named entities	`ner_person_count`, `ner_org_count`, `ner_location_count`
Style	`style_has_exclamation`, `style_question_marks`, `style_bullet_count`, `emoji_count`
Topics	`topic_tech`, `topic_business`, `topic_career`, `topic_leadership`
Hooks	`hook_score`, `hook_x_power_score`, `has_announcement_hook`
Narrative	`has_personal_story`, `has_vulnerability`, `has_contrast`, `has_adversity_learning`
Length/structure	`sentence_count`, `length_score`, `url_count`, `hashtag_count_extracted`

4.2 What Is Excluded

Follower count, log-transformed follower count, and follower tier are **not passed to any model**. Within a tier, follower count has near-zero variance — all creators in the micro tier have fewer than 10,000 followers, so the variable provides no discriminative information within that tier. More importantly, including any audience-size feature would undermine the goal: classifying content quality, not audience scale.

All engagement-derived columns (`reactions`, `comments`, `engagement_rate`, `reactions_per_word`, etc.) and aggregated author-history features (`influencer_avg_reactions`, `influencer_post_count`, etc.) are also excluded to prevent target leakage.

---

5. Modelling Approach

5.1 Algorithm Suite

Three tree-ensemble classifiers are trained per tier:

Algorithm	Key Parameters	Class Balance Handling
Random Forest	n_est=200, max_depth=8, min_samples_split=10, min_leaf=5	`class_weight="balanced"`
XGBoost	n_est=200, max_depth=4, lr=0.05, min_child_weight=5, sub=0.8	`sample_weight` from `compute_sample_weight("balanced")`
LightGBM	n_est=200, max_depth=4, lr=0.05, num_leaves=15, min_child=10	`class_weight="balanced"`

Evaluating three algorithms per tier rather than committing to a single one allows the best model to be selected independently for each segment. This is appropriate because the optimal inductive bias may differ across tiers: gradient boosting often works well on larger samples with more complex interactions; random forests can be more stable on smaller samples where overfitting risk is higher.

5.2 Why No Hyperparameter Tuning

Hyperparameter search via cross-validation requires that each training fold contain enough samples to produce stable estimates. With the medium tier at 76 training samples and the large tier at 89, a standard 5-fold CV yields approximately 15–18 samples per fold — far too few for reliable gradient estimates. Tuning only the larger tiers would introduce methodological inconsistency across tiers.

Instead, deliberately conservative fixed parameters are used throughout: shallow maximum depths (4–8), high minimum-child-weight constraints, and moderate subsampling. These choices reduce overfitting risk on small samples at the cost of potentially leaving performance on the table for the larger micro and small tiers.

5.3 Label Construction and Leakage Prevention

For each tier:

1. The training rows belonging to that tier are identified using `follower_tier` (which is computed for splitting purposes only, never passed as a feature).

2. The training-subset median engagement rate is computed — this is the class boundary.
3. Posts in both the training and test sets are labelled using this training-derived median: posts at or above it are Class 1 (Above), posts below it are Class 0 (Below).

By deriving the median from the training subset only, no information from the test distribution influences the class assignment. The test median is never computed or used.

---

6. Results

6.1 Per-Tier Performance

Tier	RF F1	XGB F1	LGBM F1	Best Model	Best F1
micro	0.6103	0.7011	0.7119*	LightGBM	0.7119
small	0.6333	0.7317*	0.5586	XGBoost	0.7317
medium	0.6761*	0.6761*	0.6000	RF / XGB	0.6761
large	0.7597*	0.6667	0.6111	RandomForest	0.7597

** = best or tied-best in tier*

6.2 Aggregate Performance

Metric	Value
Weighted average Macro F1	**0.7199**
Random baseline	0.5000
Lift over random	**+0.2199**
Models trained	12 (4 tiers × 3 algorithms)
Feature set	71 content features

The weighted F1 is computed by weighting each tier's best-model F1 by its test-set sample size (77 + 41 + 16 + 21 = 155 total test posts). This correctly weights the micro and small tiers — which have larger and more reliable test samples — more heavily than medium and large.

6.3 Confusion Matrices

Each confusion matrix shows the best-performing model for that tier.

micro (<10k) — LightGBM — F1 = 0.7119 — (n = 77)

	Pred: Below (<22.2)	Pred: Above (≥22.2)
Actual: Below (<22.2)	31	12
Actual: Above (≥22.2)	10	24

Class	Precision	Recall	F1
Below (<22.2)	0.756	0.721	0.738
Above (≥22.2)	0.667	0.706	0.686

The model is slightly stronger at predicting below-average posts than above-average ones. Precision for the below class (0.756) exceeds precision for the above class (0.667), suggesting the model occasionally over-predicts above-average outcomes.

small (10k–50k) — XGBoost — F1 = 0.7317 — (n = 41)

	Pred: Below (❤️.3)	Pred: Above (≥3.3)
Actual: Below (❤️.3)	15	7
Actual: Above (≥3.3)	4	15

Class	Precision	Recall	F1
Below (❤️.3)	0.789	0.682	0.732
Above (≥3.3)	0.682	0.789	0.732

The small-tier XGBoost model produces a striking symmetry: both classes achieve identical F1 (0.732), but with inverted precision-recall profiles. The model is more precise for below (0.789) but more sensitive for above (recall 0.789). The practical implication: when the model predicts a post will underperform, it is right 79% of the time; when a post is actually above-average, the model catches it 79% of the time.

medium (50k–200k) — RandomForest — F1 = 0.6761 — (n = 16)

	Pred: Below (<1.7)	Pred: Above (≥1.7)
Actual: Below (<1.7)	7	2
Actual: Above (≥1.7)	3	4

Class	Precision	Recall	F1
Below (<1.7)	0.700	0.778	0.737
Above (≥1.7)	0.667	0.571	0.615

With only 16 test samples, each individual prediction shifts F1 substantially — a single misclassification moves F1 by approximately 0.06. The directional finding is that the model predicts below-average posts more reliably than above-average ones, consistent with the asymmetry observed in the larger tiers.

large (>200k) — RandomForest — F1 = 0.7597 — (n = 21)

	Pred: Below (<0.3)	Pred: Above (≥0.3)
Actual: Below (<0.3)	7	2

	Pred: Below (<0.3)	Pred: Above (≥0.3)
Actual: Above (≥0.3)	3	9

Class	Precision	Recall	F1
Below (<0.3)	0.700	0.778	0.737
Above (≥0.3)	0.818	0.750	0.783

The large tier inverts the usual pattern: precision for the above class (0.818) exceeds precision for the below class (0.700). The large-tier RF model is better at confirming above-average posts than catching below-average ones — suggesting that the content signals for top-performing large-account posts are more distinctive than the signals for underperforming ones.

6.4 Feature Importance — Top 5 per Tier (XGBoost)

Feature importances are reported from XGBoost across all tiers for consistency (XGBoost was not always the best model, but is used here to enable direct comparison of feature importance without algorithm effects).

micro (<10k)

Rank	Feature	Importance
1	`style_has_exclamation`	0.047
2	`has_direct_address`	0.038
3	`style_exclamation_marks`	0.034
4	`url_count`	0.033
5	`style_has_question`	0.033

small (10k–50k)

Rank	Feature	Importance
1	`topic_career`	0.059
2	`style_has_exclamation`	0.057
3	`has_contrast`	0.048
4	`ner_location_count`	0.035
5	`has_vulnerability`	0.033

medium (50k–200k)

Rank	Feature	Importance
1	`emoji_count`	0.097

Rank	Feature	Importance
2	`style_has_emoji`	0.065
3	`unique_emoji_count`	0.062
4	`topic_count`	0.057
5	`text_difficult_words_count`	0.047

large (>200k)

Rank	Feature	Importance
1	`text_difficult_words_count`	0.114
2	`url_count`	0.091
3	`readability_flesch_kincaid`	0.068
4	`length_score`	0.059
5	`readability_gunning_fog`	0.054

7. Analysis & Discussion

7.1 All Tiers Beat Random

The most fundamental result is that every tier's best model achieves F1 substantially above 0.500. The improvements range from +0.176 for medium to +0.260 for large. This is not a trivial finding: the features are content-only — no information about the creator's audience size, posting history, or platform signals is included. The result confirms that content structure, style, and topic choice carry genuine predictive signal for engagement outcomes within every creator segment.

7.2 Interpreting the Weighted Aggregate

The weighted average Macro F1 of 0.7199 should be interpreted carefully. Because micro (n_test=77) and small (n_test=41) account for 75% of the total test sample, they dominate the weighted average. The aggregate result is primarily a statement about how well content features predict engagement for accounts with fewer than 50,000 followers — which is where most LinkedIn creators sit.

The medium and large tiers contribute to the weighted average, but their individual F1 estimates carry high variance. With 16 and 21 test samples respectively, a shift of two predictions changes F1 by approximately 0.06–0.08. For these tiers, the direction of the result (F1 > 0.500, genuine signal exists) is more informative than the precise value.

7.3 Model Selection Pattern

No single algorithm dominates across all tiers:

- **micro:** LightGBM (0.7119) beats XGBoost (0.7011) and RF (0.6103)
- **small:** XGBoost (0.7317) beats RF (0.6333) and LightGBM (0.5586)

- **medium:** RF and XGBoost tie (0.6761); LightGBM lags (0.600)
- **large:** RF (0.7597) substantially outperforms XGBoost (0.6667) and LightGBM (0.6111)

The pattern suggests that gradient-boosted methods (LightGBM, XGBoost) are better suited to the larger samples of micro and small tiers, where they have enough data to learn complex interactions. Random Forest's stronger performance in medium and large tiers is consistent with its known stability advantage under small sample conditions — RF's ensemble of independent trees is more robust to limited training data than sequential boosting methods.

The notable underperformance of LightGBM on small (F1=0.5586) and large (F1=0.6111) is also consistent with LightGBM's leaf-wise growth strategy, which can lead to overfitting on small datasets if not carefully regularised. Despite the regularisation applied through `num_leaves=15` and `min_child_samples=10`, the small and large samples appear insufficient for LightGBM's growth pattern to converge well.

7.4 Feature Divergence Across Tiers

The most interpretively rich finding is how dramatically feature importance shifts across creator segments. The four tiers are not just different sizes of the same phenomenon — they represent fundamentally different content-engagement dynamics.

Micro creators: conversational style signals. The top features for micro creators are almost entirely stylistic and conversational — exclamation marks (flag and count), direct address, question presence, and URL count. Micro-creator audiences are small and personal; the creator-follower relationship is closer to a one-to-one conversation than a broadcast. Content that mimics the energy and directness of personal communication (enthusiasm through exclamation, explicit audience address, rhetorical questions) outperforms more formally structured posts. The negative signal of URL count is also consistent: micro audiences respond to native, personal content and penalise posts that feel like they are redirecting attention elsewhere.

Small creators: career-focused authentic storytelling. Small creators show a decisive shift toward semantic content signals. `topic_career` is the top feature by a meaningful margin (0.059), followed by `style_has_exclamation` as the only stylistic holdover from micro. The appearance of `has_contrast` (0.048) and `has_vulnerability` (0.033) reveals that narrative authenticity — posts that acknowledge difficulty, challenge, or personal setback — is the dominant engagement driver for small-tier accounts. These creators have established enough credibility to have an audience, but their audience still expects genuine personal insight rather than polished corporate messaging. The presence of `ner_location_count` is intriguing: specificity of location (mentioning real places, events, or geographies) appears to strengthen engagement, possibly by adding authenticity and concreteness.

Medium creators: visual and multimedia signals. The medium tier undergoes a structural break: the top three features are all emoji-related (`emoji_count`, `style_has_emoji`, `unique_emoji_count`), with no overlap with the conversational style signals that dominate micro and small. At 50,000–200,000 followers, creators are posting into a feed with significant competing content. Visual cues — emoji as structural markers and attention anchors — may serve to interrupt scanning behaviour and draw readers into the post. The appearance of `topic_count` (0.057) suggests that breadth of topic coverage also matters at this scale, possibly reflecting that larger audiences benefit from content that addresses multiple angles rather than a single focused point.

Large creators: text complexity and readability. Large accounts show the clearest divergence of all. Three of the top five features are readability metrics: `text_difficult_words_count` (0.114), `readability_flesch_kincaid` (0.068), `readability_gunning_fog` (0.054). The direction of these effects requires careful interpretation — harder words and higher FK grade are associated with above-average engagement for large accounts. This is counter-intuitive if readability is assumed to mean "simpler is better," but for large LinkedIn accounts (executives, senior professionals, thought leaders), the audience demographic is itself highly educated and professionally experienced. Sophisticated vocabulary signals expertise and authority rather than obscurity. `url_count` as the second feature (0.091) also inverts the micro pattern — for large accounts, links may signal comprehensive sourcing and credibility rather than platform-exit behaviour.

The degree of divergence across these four feature profiles strongly supports the per-tier modelling choice. A single model would be forced to learn a compromise between these four different engagement dynamics, suppressing tier-specific signal. Training separate models allows each tier's features to receive full weight without being averaged away.

7.5 Confusion Matrix Patterns

A consistent asymmetry appears across most tiers: the model identifies below-average posts with higher precision than above-average posts, but above-average posts with higher recall. In practical terms, when the model predicts "this post will underperform," it is more likely to be right than when it predicts "this post will over-perform." But when a post genuinely over-performs, the model tends to catch it.

The exception is the large tier, where the above class achieves both higher precision (0.818) and reasonable recall (0.750). For large accounts, the characteristics of above-average posts appear particularly distinctive — they are fewer in number but more consistently recognisable by the content features.

This asymmetry has a practical implication: the model is more useful as a negative filter (flagging content likely to underperform) than as a positive identifier (certifying content as likely to over-perform). False positives for the above class — posts predicted to succeed that do not — are more common than false negatives.

7.6 Sample Size as the Binding Constraint

Medium (76 train, 16 test) and large (89 train, 21 test) represent a fundamental limitation. Even if the content features carry real predictive signal for these segments, the models cannot fully surface that signal with so little data. Conservative parameters help prevent overfitting but also limit the model's capacity to learn higher-order interactions.

The micro and small results (combined 452 training samples, 118 test samples) are the most reliable in the experiment. Any conclusions about the viability of content-only per-tier classification should rest primarily on these two tiers.

---

8. Conclusions

This experiment demonstrates that content features carry genuine predictive signal for engagement outcomes within each creator segment when posts are evaluated against their own tier's median. Four conclusions stand out:

1. Content is predictive within every segment. All four tiers' best models achieve F1 above 0.500, with lifts ranging from +0.176 to +0.260. The experiment confirms that content structure and style — not just creator size — drive within-segment engagement variation.

2. Feature signals are tier-specific. The divergence in feature importance across tiers is the most important substantive finding. Micro creators respond to conversational energy; small creators to career-focused authentic storytelling; medium creators to visual and multimedia cues; large creators to sophisticated, complex text. This is not noise — it reflects the distinct audience relationships and content expectations that define each segment. A content scoring system that applies a single feature-weighting scheme to all creators is, at minimum, a poor approximation of these underlying dynamics.

3. The per-tier approach is the correct fairness frame. By training on within-tier labels and evaluating posts against their own segment's median, each creator is judged against peers rather than against the full distribution. This removes the structural advantage of high-engagement tiers and the structural disadvantage of low-engagement tiers from the classification outcome.

4. Sample size is the binding constraint for medium and large. The experiment is fundamentally limited by dataset size for the two largest tiers. The results for medium (n_test=16) and large (n_test=21) are directional estimates, not stable performance measurements. Expanding the dataset — particularly for established-creator posts — is the highest-leverage improvement available to this experiment design.

Practical implications. For a content recommendation or scoring system aimed at helping LinkedIn creators, this experiment provides a viable foundation: separate scoring models per tier, trained on content-only features, each calibrated to its segment's own median. The model produces actionable directional predictions (likely to over- or under-perform relative to peers) without relying on audience size as a shortcut. The feature importance results also provide the basis for tier-specific content guidance: micro creators benefit from conversational style adjustments, small creators from narrative authenticity, medium creators from visual structure, and large creators from substantive complexity.