

Success Predictor for LinkedIn Posts

Transfer Project
CAS ML 2025 @ HSLU
Reto Lämmler
Zurich, 3.2.2025



Sergio P. Ermotti · 3rd



Group CEO and President of the Executive Board of UBS AG

Zurich, Zurich, Switzerland · [Contact info](#)

307,457 followers

Followed by Yvan, Helmuth and 264 others you know

Following

Connect

...

65% DROP IN REACH? |

LINKEDIN'S 2025

LinkedIn silently changed its algorithm, and the old strategies just don't work anymore.

Sonu Kushwaha

Goal of Transfer Project

1. Train a **model to predict success** for LinkedIn Posts
2. Understand **what features drive success** and what punishes you
3. Compare impact of features **before vs. after the algorithm change** in 2025

Data Acquisition



Valentin Binnendijk and 77 others



158 comments • 4 reposts

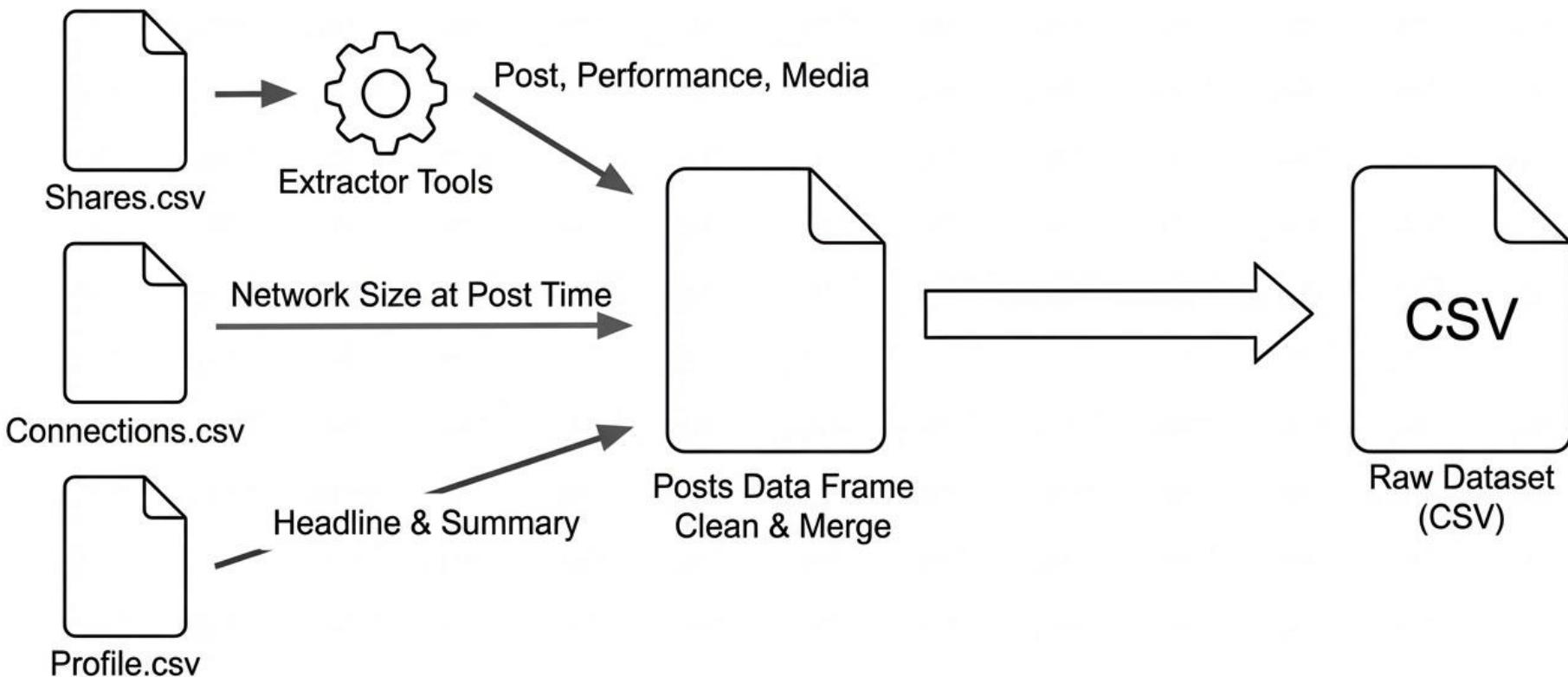
Reactions



A screenshot of a LinkedIn page titled "Download my data". The page includes a back button, a title, and a main section with two options: "Download larger data archive" (selected) and "Want something in particular? Select the data files you're most interested in." Below these are checkboxes for "Articles", "Invitations", "Profile", "Recommendations", and "Registration". A large blue "Request archive" button is at the bottom. Below the button, text says "Your download will be ready in about 24 hours". At the very bottom, there's a note about Member Data Portability APIs and a link to the Help Center.

**10k posts
from 42 people**

Data Preprocessing



Feature Engineering

Post Content Features:

- Post Content Length
- Hook Length
- Linebreak Count
- Emoji Count
- Hashtag Count
- Link Count

Posting Time Features:

- Hour of day (0...23)
- Day of week (0...6)

NLP Features:

- Sentiment Score
 - *nlptown/bert-base-multilingual-uncased-sentiment*
- Semantic Alignment
 - *paraphrase-multilingual-MiniLM-L12-v2*
 - cosine similarity Profile Summary <> Post Content



Reto Laemmli · You
UX-driven founder, exited CEO (TestingTime), Curious Learner
1mo · Edited ·

I need your help! 😊

I'm back in school for a CAS in Machine Learning (ML) at [Lucerne University of Applied Sciences and Arts](#). 🤓

For my thesis project, I'm building a machine learning model to quantify (and predict) why some LinkedIn posts take off while others disappear, especially under LinkedIn's new LLM-driven algorithm.

Short intro on the new algorithm: <https://lnkd.in/eVsxEQuR>
Deep dive: <https://lnkd.in/e8xuPCev>

What I need from you?

I need some of your public LinkedIn data, which you can download easily from your account: your posts, profile basics, connections

Otherwise: No data → no model → no graduation 😊

What do I offer in return

1. Early, personalized insights into when and why your posts perform.
2. If this becomes a product → free lifetime access 🤩

Please comment "DATA" and I'll send you simple instructions on how to download and share it with me safely.

P.S. I never believed to ever apply this "comment" hack. Godspeed to the algorithm to spread my good word! 😊

P.P.S. Left image is Nano Banana. Funny how my Oura ring changed the finger 😊



Valentin Binnendijk and 77 others

158 comments · 4 reposts

Reactions



Target variable

Weighted Relative Engagement Score:

$$S_{rel} = ((Likes * 1 + Comments * 3) / Network Size) * 1000$$

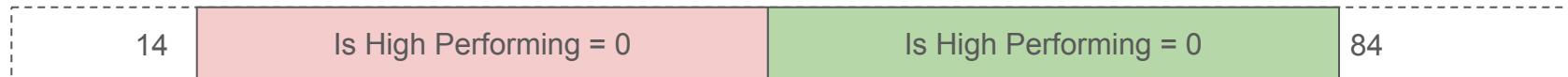
Binary Classifier:

Is High Performing (0 / 1)

Mapping Engagement Score to “Is High Performing”

User A:

Median = 49



User B:

Median = 35



Why Median per user?

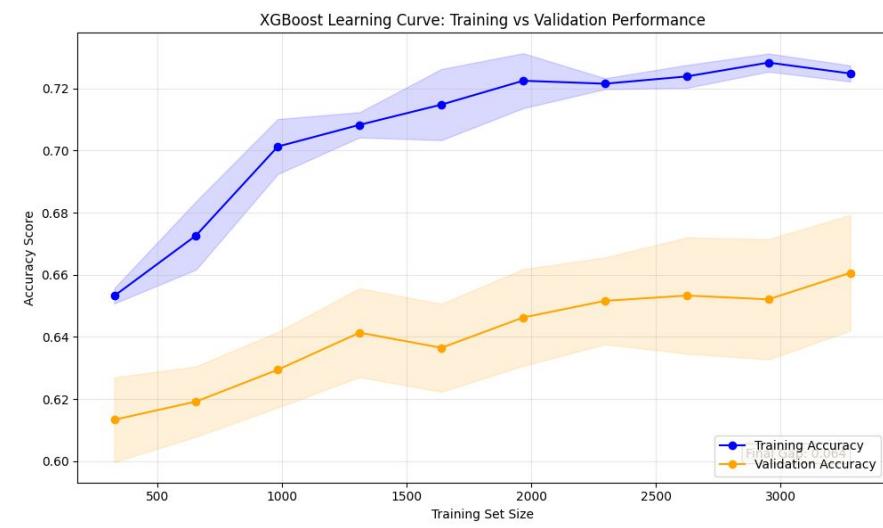
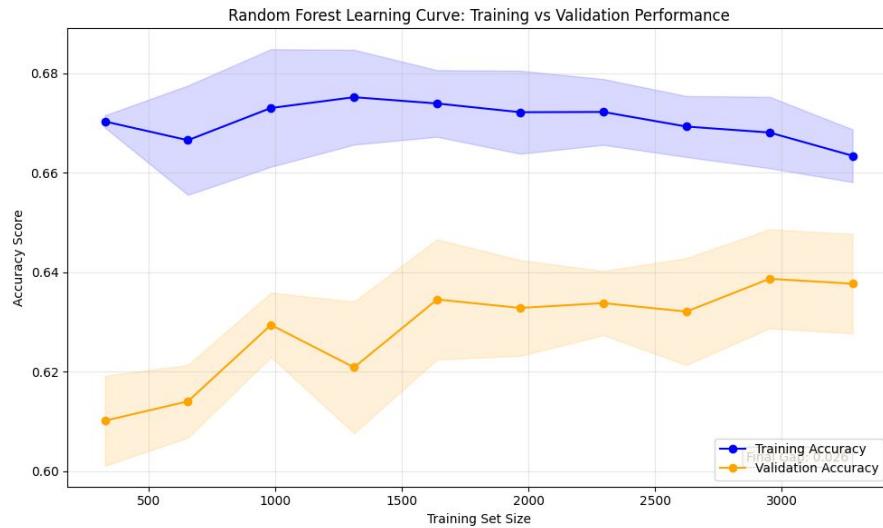
- 50/50 class balance
- Outlier robustness
- Equalize performance
- Mitigating identity bias

Training Set Balancing

- **User Capping:** A maximum limit of **350 posts per user** was enforced.
- **Outlier Management:** The **Relative Engagement Score** was capped at **150**
- **Min. Content Length:** Content / reposts below **50 characters** were removed.

After the balancing, **5124 posts** were left for training.

Learning Curve: Random Forest vs. XGBoost



Hyperparameter Tuning

| 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 |

Random Forest:

Tuning reduces overfitting and stabilizes predictions on noisy social-media data.

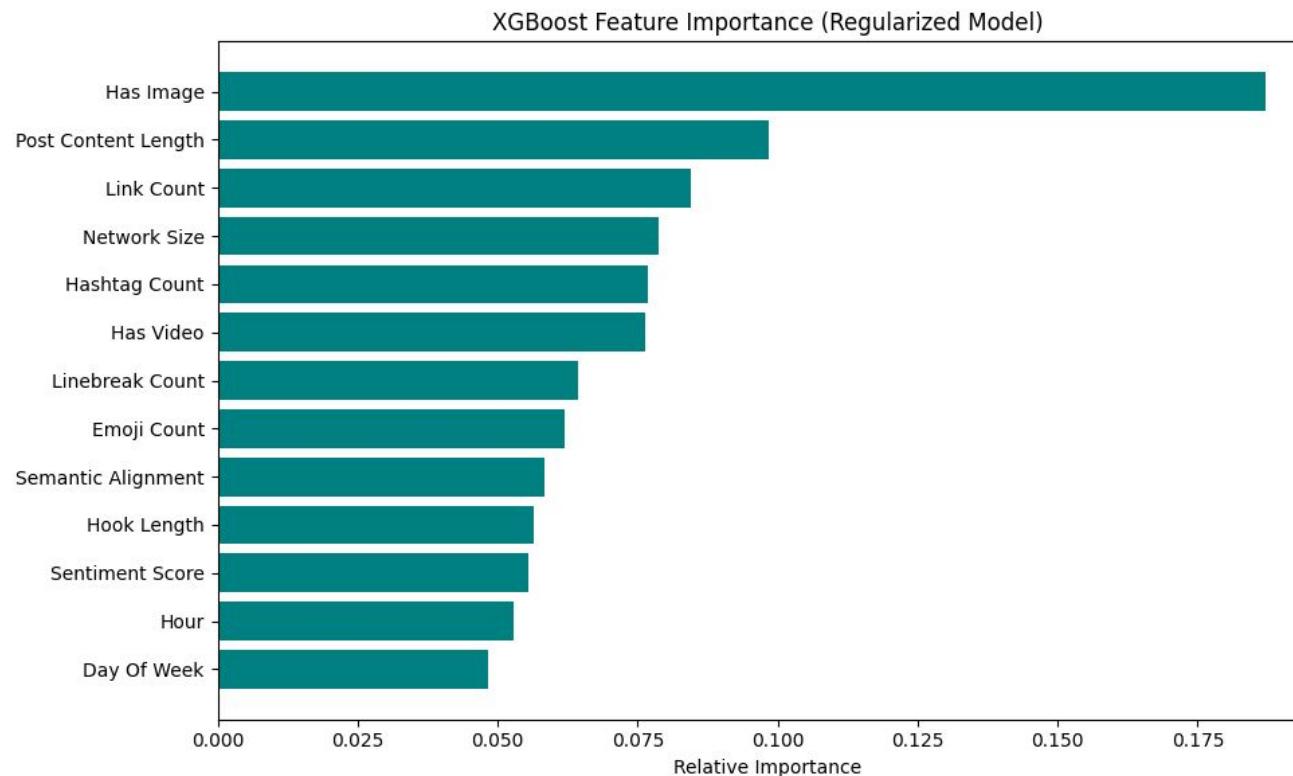
XGBoost:

Tuning forces the model to learn slowly and robustly, improving generalization.

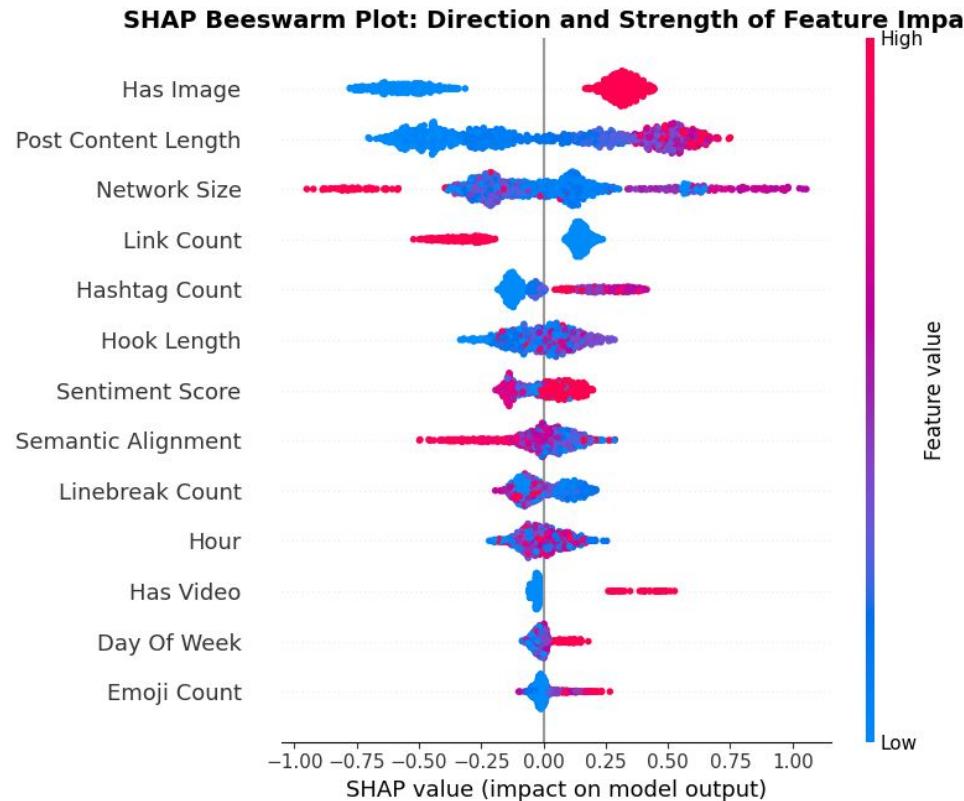
Test Set Performance Comparison

| 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% |

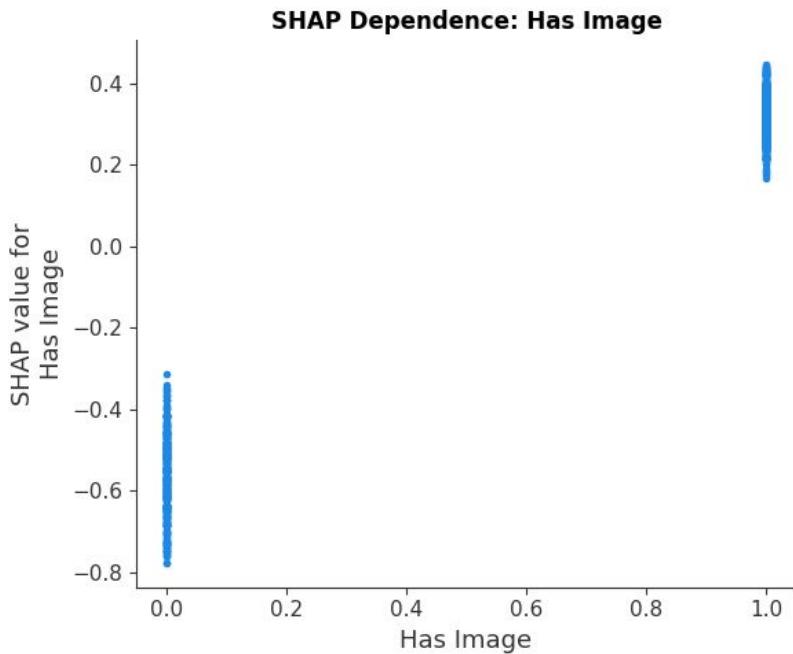
Feature Importance (XGBoost)



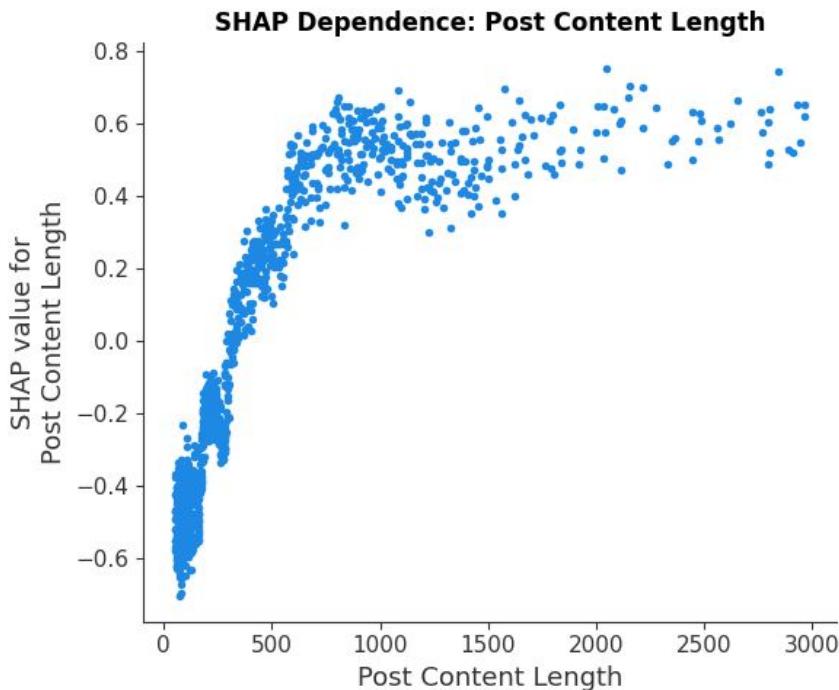
Global SHAP Beeswarm



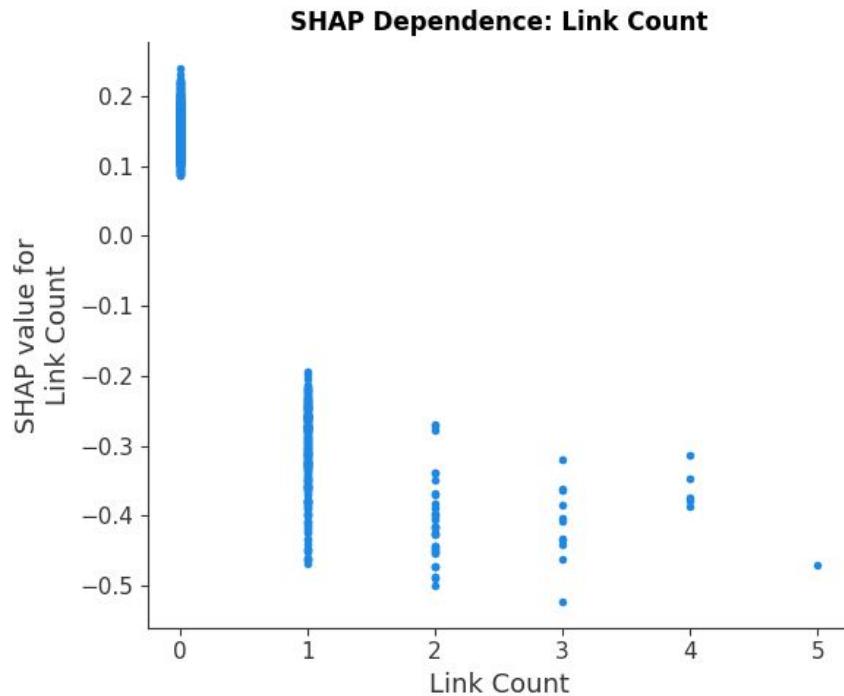
An image is a must have



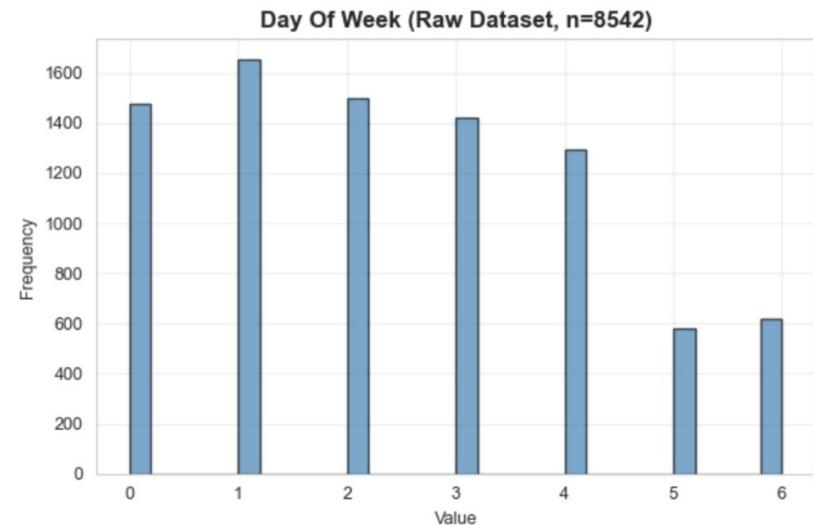
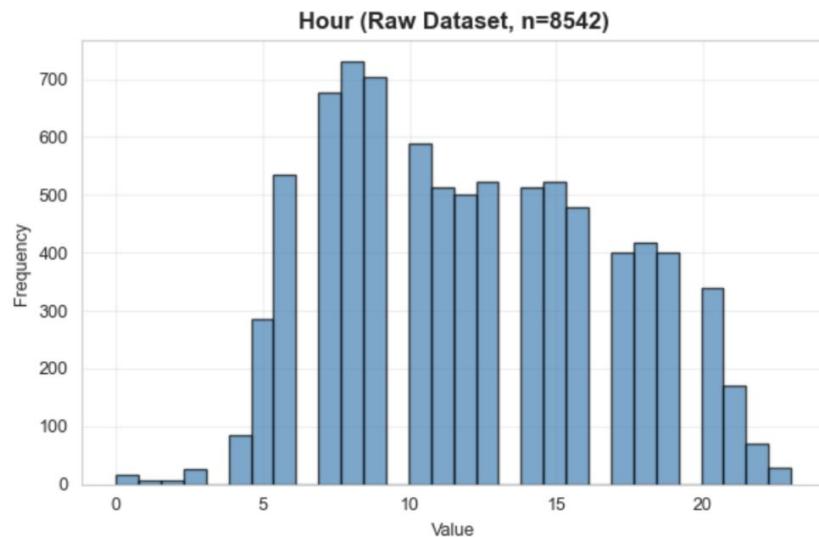
Short content gets penalized



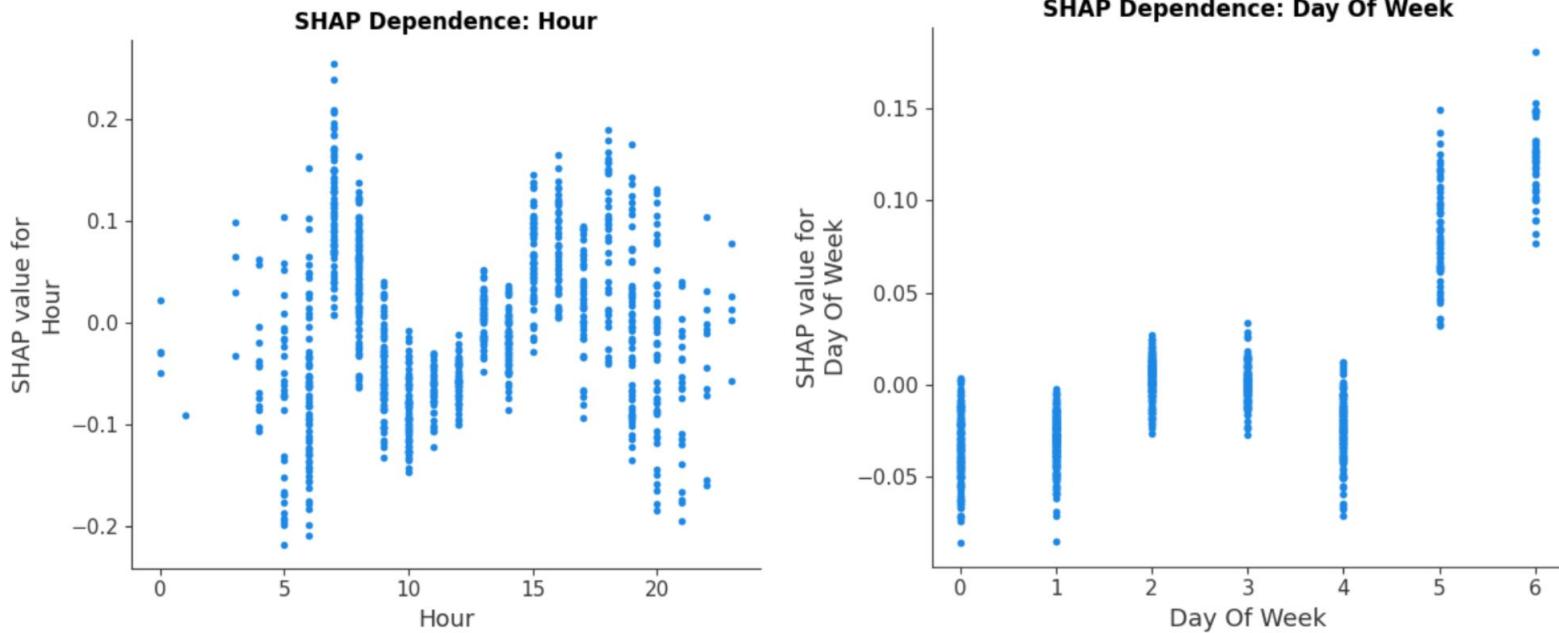
External links in your post are penalized



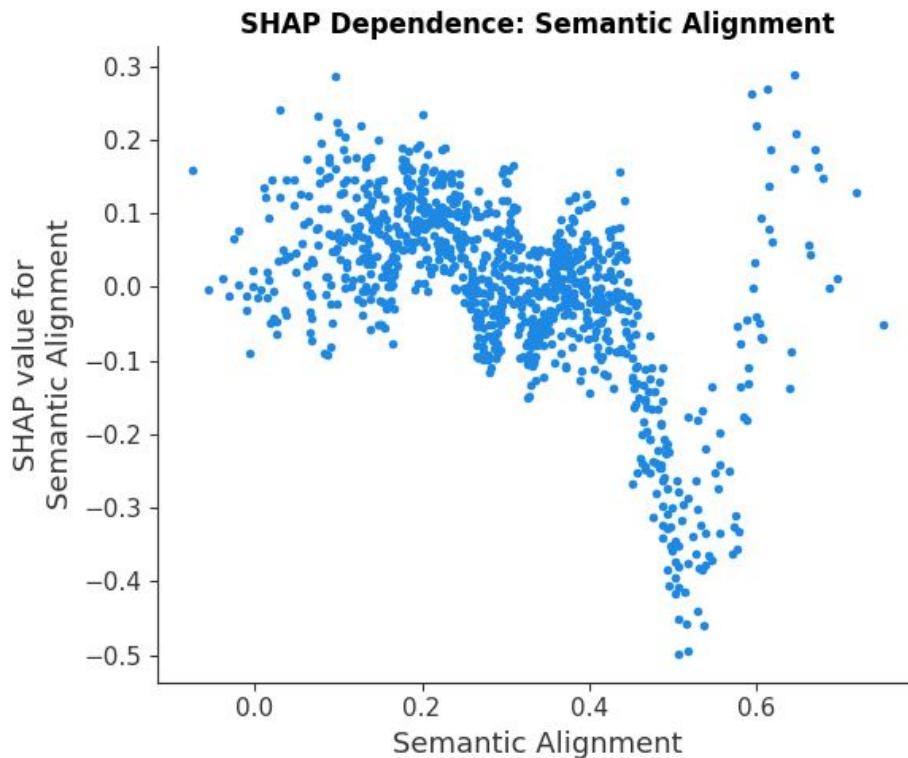
The dataset believes the myth of Tuesday morning



But posting time doesn't matter so much
(maybe try Saturday or Sunday)

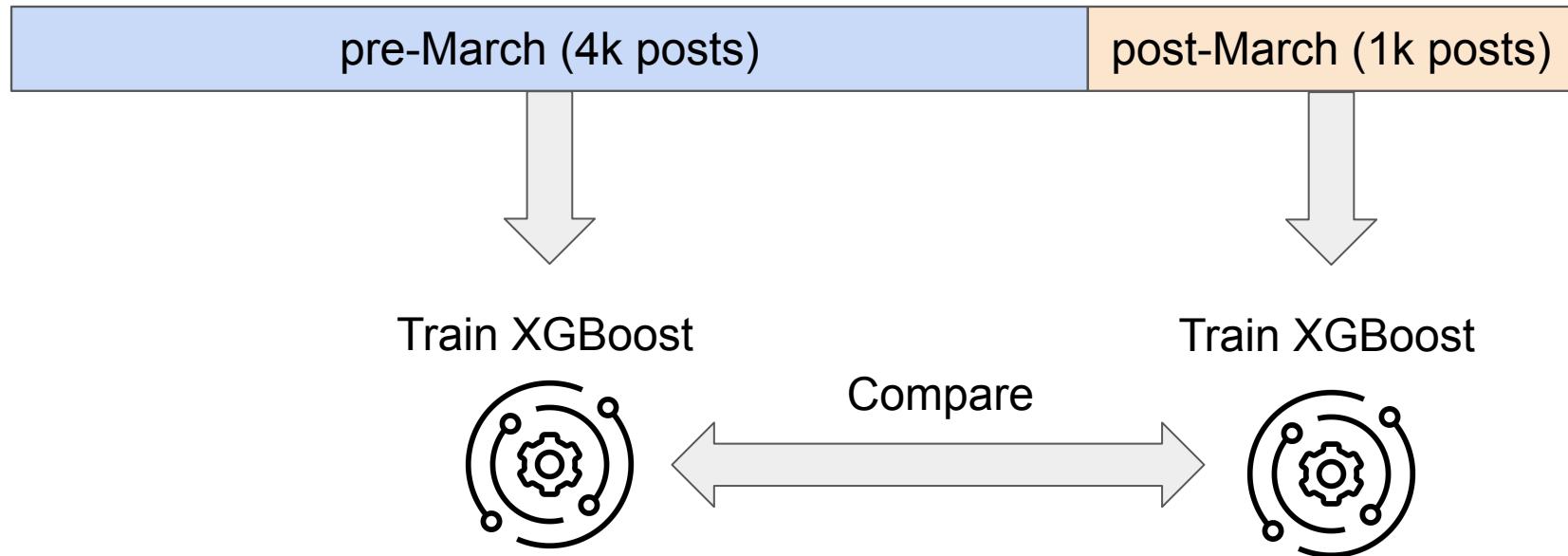


Semantic alignment has ambiguous influence

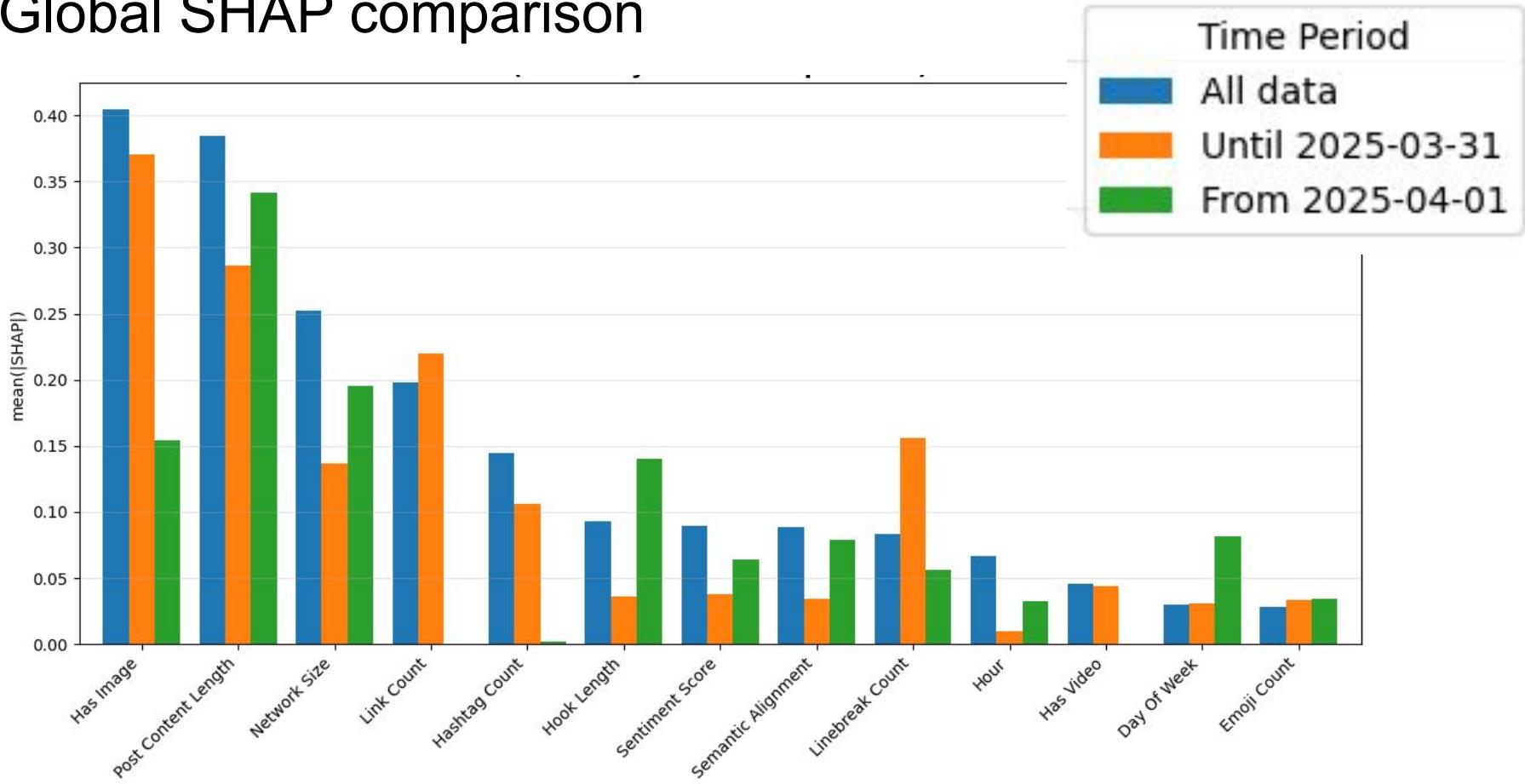


Concept Drift: LinkedIn algorithm change in 2025

Split dataset at March 2025:



Global SHAP comparison



Conclusion

- **Network Size** and **Semantic Alignment** increase in importance
- **Content and Hook Length** remain important drivers
- **Day of week** strongly increases, while **Link Count** becomes completely unimportant

Remark: Due to small post-March dataset, the results require caution

Future Outlook

- **Retrain with more data** post-March 2025 and outside the author's network
- Apply **GroupKFold** to test model generalization to unseen users
- Deploy model as **Mini-App** with recommendations for post improvements

Recommendations for your next LinkedIn post

1. **Min. 500 characters** long. Avoid simple / empty reposts.
2. Add at least one **image (or video)** to your post
3. **Avoid external links** in the post. Add them in the comments.

In general: increase your network size with meaningful connections

Thank you!



Preprocessed Features

- **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+1 from 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