Detecting Entrepreneurial Behavior in the #sahm TikTok Ecosystem Using Hashtag-Based Machine Learning

Ashley Cong, Class of 2026

Wellesley College, Data Science Major Capstone

Introduction

TikTok is one of the fastest growing social media platforms known for its short-form videos and fast content circulation. Visibility on TikTok is shaped by engagement signals and hashtag-driven content categorization, so creators use them to signal identity, categorize content, and reach audiences. Hashtag patterns provide a scalable, interpretable way to detect entrepreneurial behavior across millions of posts.

The #sahm (Stay-at-home-mom) community on TikTok represents a distinct subculture of creators who share daily routines, parenting advice, and glimpses of domestic life. These creators form a rapidly growing niche on TikTok, where the boundaries between domestic labor, emotional labor, and entrepreneurial labor blur. Many creators leverage family-centered content to promote small businesses, digital products, or coaching services, embodying an entrepreneurial motherhood.

This project develops a computational framework to identify business-oriented stay-at-home-mom creators using hashtag-based signals, text-embeddings, and machine-learning models. This approach provides a scalable method for identifying entrepreneurial behaviors within the #sahm creator ecosystem.

Research Question

How can we identify business-oriented and entrepreneurial #sahm creators on TikTok using hashtag-based features, and which modeling approach provides the best predictive accuracy?

Data

This project draws on U.S. based TikTok posts associated with #sahm creators collected from the TikTok API and includes posts from 2018-2024 with their associated hashtags. A user-level feature was constructed to characterize creators' content patterns and identify entrepreneurial behavior.

Data Components

- Post-level file with one row per (*user*, *post*, *hashtag*) pair.
- Top 25,000 hashtags used by ≥30 creators to build the model vocabulary
- Aggregated file listing each creator's total post count
- Manually labeled set of 60 creators

Data Processing and Feature Construction

- Standardized creator identifiers and cleaned hashtag text.
- Built a sparse user x hashtag count matrix using only high-frequency hashtags to reduce noise and file size
- Converted raw counts into TF-IDF vectors representing each creator's topical profile
- Connected manually labeled creators to their TF-IDF representations to form an evaluation subset.

Fig 1. Distribution of Business Hashtag Usage Across Creators

Exploratory Data Analysis

Since raw TikTok data is noisy and computationally heavy, the analysis centers on user-level hashtag behavior, which provides consistent signals across heterogeneous posting styles. Initial exploration showed highly skewed hashtag usage. Business-related hashtags occur relatively infrequent overall and cluster within a subset of creators. These hashtags tend to co-appear with commerce-related tags that tend not to appear among high-frequency lifestyle hashtags. The distribution suggests that entrepreneurial creators can be identified through specific, intentional hashtags.

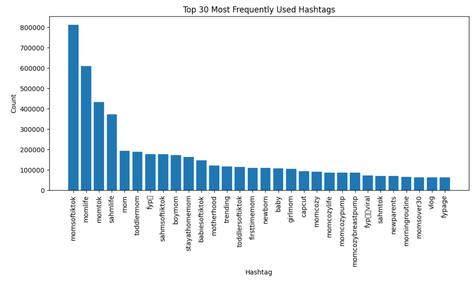

Fig 2. Top Hashtags Frequency Plot

Methodology

A user x hashtag sparse matrix was constructed with the top 25,000 hashtags that appear across at least 30 creators. This vocabulary restriction reduces noise and preserves interpretable patterns. TF-IDF representations for each creator is then constructed to capture relative emphasis on topics. A business-hashtag proportion was created based on a curated set of hashtags as an interpretable rule-based indicator.

Approach 1: Unsupervised Clustering

KMeans clustering (k=2) was applied to the TF-IDF matrix to explore whether entrepreneurial creators naturally forma a distinct content cluster. Cluster identities are inferred by comparing composition against manually labeled creators.

Approach 2: Rule-Based Classification

A heuristic classifier was developed using each creator's business-hashtag proportion. Decision threshold is tuned on the manually labeled subset to maximize F1 score.

Approach 3: Supervised Machine Learning

A logistic regression classifier is trained on TF-IDF vectors from manually labeled data using stratified train-test splits and 5-fold cross-validation. A k-nearest neighbors (KNN) baseline was also evaluated to compare the performance of non-parametric models.

Approach 4: Transformer-Based Text Embeddings

Video descriptions were aggregated at the user level and encoded using a lightweight sentence-transformer model (MiniLm-L6-v2) to generate dense text embeddings. A logistic regression classifier was then trained on these embeddings to evaluate whether semantic signals captured through captions improve identification of entrepreneurial creators.

Results

The evaluation compares three approaches using a manually labeled set of creators with hashtag data to be represented in the TF-IDF matrix. The results show that no single method dominates across all metrics.

	F1	Precision	Recall	ROC AUC
Clustering	0.67	0.55	0.86	N/A
Rule-Based	0.75	0.67	0.86	0.81
TF-IDF	0.33	0.50	0.25	0.83
Transformer	0.67	0.71	0.63	0.88

The rule-based method achieved the highest F1 score, which shows its ability to identify creators who consistently deploy explicit hashtags. However, the transformer-based model achieved a balanced performance overall.

1. Comparative Accuracy Across Methods
Clustering showed high recall but low precision, showing that entrepreneurial creators are not condensed in a single cluster but rather diffused across broader lifestyle. The rule-based classifier's precision and recall drop when business creators' content are more nuanced. TF-IDF logistic regression model struggled to distinguish entrepreneurial identity based solely on hashtag patterns, but cross-validation provides more consistent behavior when applied at scale.

2. Patterns of Disagreement:

- Creators flagged only by the transformer-based model use descriptive language about selling or promoting without marking these posts with business-oriented hashtags.
- Creators flagged only by the rule-based method tend to use a small set of explicit business hashtags but may not have diverse business-oriented content.

3. Structural Insights Across Representations
Dimensionality reduction of hashtag-based and text-based feature spaces shows that entrepreneurial creators do not form a bounded group but instead overlap regions within broader #sahm ecosystem. All approaches show that creators form diffuse pockets rather than discrete clusters, indicating that entrepreneurial identity is blended with lifestyle content rather than a standalone niche. The structural pattern explains why unsupervised methods struggle and why supervised models that incorporate semantic cues are can better detect entrepreneurial identity by leveraging subtle contextual differences rather than relying on clear separability in feature space.

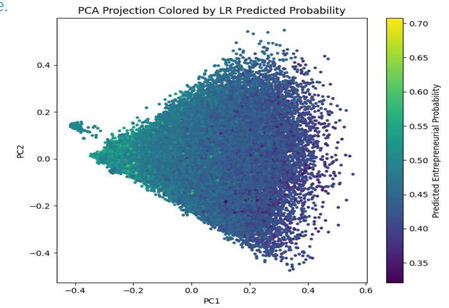

Fig 3. PCA Projection of Creator Hashtags Colored by Predicted Entrepreneurial Probability

The smooth gradient in predicted probabilities shows that entrepreneurial signals are subtle and dispersed, highlighting why logistic regression outperforms unsupervised clustering

Conclusion

This study evaluated four methods to identify entrepreneurial #sahm creators on TikTok. No single approach dominated across all metrics, reflecting the multimodal nature of entrepreneurial identity on the platform. The rule-based method achieved the highest F1 score, indicating that explicit hashtags are strong markers when present. The transformer-based model showed highest precision ad AUC by capturing semantic cues embedded in captions. In contrast, TF-IDF logistic regression underperformed on the labeled subset but showed more stable behavior under cross-validation.

The PCA projection show that entrepreneurial creators do not form a distinct separable group but rather dispersed within broader lifestyle content. This explains the limitations of unsupervised clustering and highlights the effectiveness of models that incorporate richer contextual information such as curated hashtags or semantic embeddings. Across methods, findings show that entrepreneurial identity on TikTok is context-dependent, and that hybrid approaches leveraging explicit markers and textual nuance create the best option for large-scale identification.

Discussion

The mixed performance across methods that each capture different facets of entrepreneurial behavior highlights that no single feature type is sufficient on its own. However, several limitations temper these findings. The manually labeled dataset is relatively small, which constrains the ability of supervised model to learn robust patterns. Many creators lack either sufficient hashtag variety or caption text, resulting in inconsistent feature coverage across users. Hashtag-based analyses inherently privilege creators who narrate their business identity explicitly, underrepresenting creators who signal through visual or audio interaction.

Future work can expand beyond these limits. Transformer-based caption models can be scaled using pretrained architectures or fine-tuning on #sahm-specific linguistic patterns. Multimodal models can also identify entrepreneurial cues shown visually. Active learning pipelines that identifies ambiguous business creators for human labeling can improve supervision quality. Beyond methodological improvements, this framework has broader applicability for studies of digital labor and influencer economies. It can support research on algorithmic visibility and commercialization of domestic life. Applying the model longitudinally can also allow researchers to study the career trajectories and temporal dynamics of online labor.

Acknowledgements

Special thanks to Professor Eni Mustafaraj from the Data Science Department of Wellesley College for the help and support throughout.

The GitHub code for this project has received help from GenAI.