Personal vs Promotional: Modeling Mompreneur Narratives in #SAHM TikTok

Ruth Jin Class of 2026
Wellesley College DS340H

Background & Research Question

Mothers who run home-based businesses (mompreneurs) use TikTok not only to share daily life but also to promote products and income opportunities. The #SAHM (Stay-At-Home Mom) community has become a major hub for these creators, blending personal narratives about motherhood with entrepreneurial messaging.

Prior research shows that social media narratives often mix emotional storytelling with promotional intent, shaping audience trust and engagement (Abidin, 2016). However, little work has examined how mompreneur creators specifically navigate the line between personal storytelling and promotional content, especially at scale.

Research Question: How do mompreneur creators within the #SAHM TikTok community balance personal and promotional storytelling content, and how does this balance relate to audience engagement at the creator level?

Data & Methods

Dataset

- Voice-to-Text SAHM Sub-Dataset: 269,867 posts with automatic speech transcription (approx. 10% of total #SAHM posts)
- Posts from 2021-04-15 to 2025-05-25

Original Variables

- Text: video_description (caption + hashtags), voice_to_text (automatic speech transcription)
- Engagement: like_count, comment_count, share_count
- Metadata: create_time (timestamp), id, username (anonymized)

New Variables

- text_all = video_description + voice_to_text, word count
- Engagement: like_rate, comment_rate, share_rate (per view), views
- Lexicon-based counts:
  - promo_hits (e.g., sales language, calls to action)
  - personal_hits (e.g., family, emotions, daily life)
- Continuous posterior probability:
  - p_promotional: Classifier-estimated probability that a post is promotional, derived from a TF–IDF logistic regression model

Workflow Preview

Voice_to_Text #SAHM Dataset → Identify Mompreneur Creators → Text Preprocessing (caption + speech) → Lexicon-Based Seed Labeling → TF–IDF Logistic Regression Classifier → Human Validation → Creator-Level Engagement Modeling

Mompreneur Creator Identification (seed labels) - Users who posted at least one #SAHM video containing mompreneur hashtags

Total unique users in dataset: 59149, Mompreneur users: 5510
Total posts in dataset: 269867, Mompreneur posts: 48971

Classification (Trained on 48,971 seed-labeled posts)

TF–IDF + Logistic Regression to identify promotional vs personal storytelling - 80% training, 20% validation, max_features = 50,000

Engagement Modeling

OLS regression to examine how degree of promotional storytelling relates to engagement outcomes

Exploratory Data Analysis

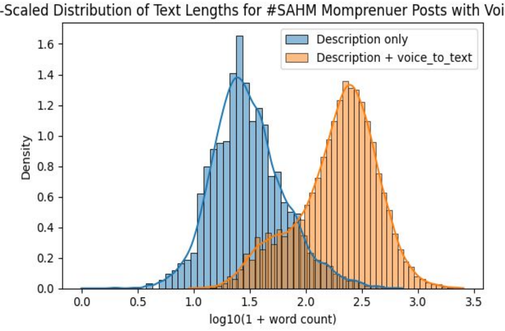

Figure 1. Distribution of Text Lengths for #SAHM Mompreneur Posts with voice_to_text. Hashtags and captions on TikTok are often optimized for algorithmic visibility rather than narrative meaning. In contrast, voice-to-text captures spoken storytelling, where creators describe personal experiences and explain products. Because mompreneur content frequently relies on spoken explanations and demonstrations, voice-to-text provides a richer and more authentic signal of storytelling style than description alone.

Classifier & Validation

High-confidence personal and promotional posts were identified using curated keyword lexicons capturing sales language (e.g., discounts, links, affiliate terms) and personal storytelling cues (e.g., family references, emotions, daily routines). Posts with strong signals for only one category were used as training data.

```
Fitting CAPTION-ONLY model...  only text_description

=== CAPTION-ONLY MODEL: validation performance ===
              precision    recall  f1-score   support

    personal     0.986     0.930     0.957      3829
  promotional    0.607     0.891     0.722       466

Fitting DESC+VT model...  text_description + voice_to_text

=== DESC + HASHTAGS + VOICE_TO_TEXT MODEL: validation performance ===
              precision    recall  f1-score   support

    personal     0.993     0.975     0.984      3829
  promotional    0.819     0.944     0.877       466
```

Table 1. Effect of Voice-to-Text on Storytelling Classification Performance. Comparison of storytelling classifier performance using caption-only text versus caption + voice-to-text. Incorporating voice-to-text substantially improves accuracy and F1 score, particularly for identifying promotional content, indicating that spoken narration captures richer promotional cues than captions alone. I proceeded with Model 2.

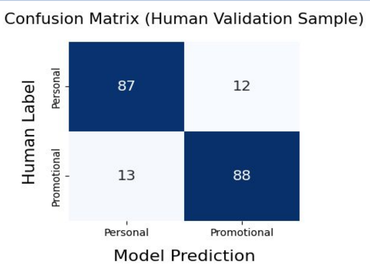

Figure 2. Human Validation Confusion Matrix for Storytelling Classifier. Confusion matrix comparing human-annotated labels and model predictions for 200 posts, sampled evenly from posts classified as personal and promotional. Overall agreement is high (accuracy = **87.5%**, Cohen's κ = 0.75), with most errors occurring in mixed or borderline storytelling cases.

Results

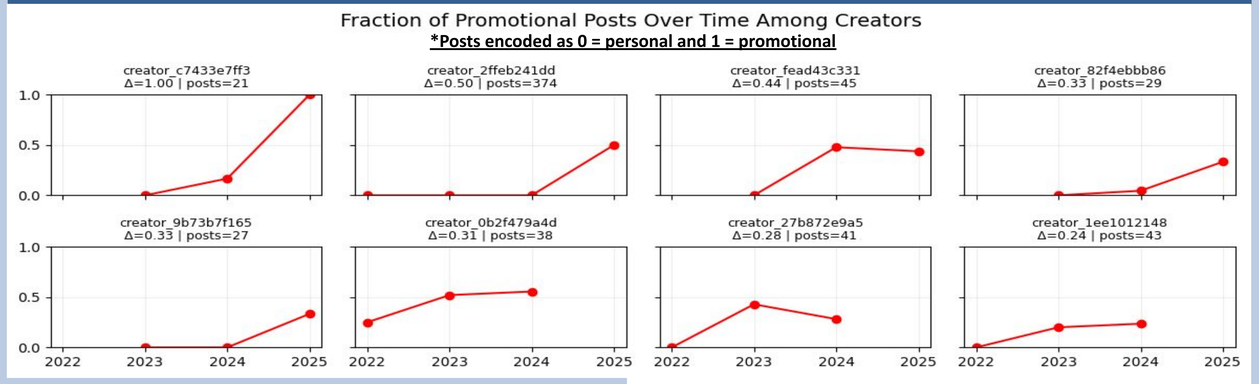

Figure 4. Creator-level trajectories of promotional storytelling over time. This figure shows year-by-year changes in mean promotional probability for eight mompreneur creators. Across creators, promotional storytelling generally increases over time. This pattern is consistent with an audience-building trajectory, in which creators initially share personal experiences of motherhood to establish authenticity and a stable follower base, and subsequently leverage that audience to promote products or income opportunities. While individual trajectories vary, the overall upward trend suggests that promotional content often emerges as a later phase of creator strategy rather than a dominant mode from the outset.

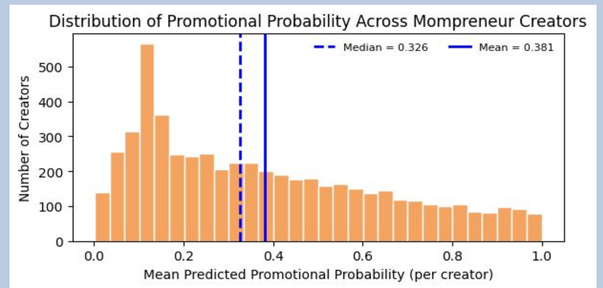

Figure 3. Distribution of creator promotional probability across creators. Histogram of mean predicted promotional probability across mompreneur creators (**n = 5,510**). Each value represents the average promotional likelihood of a creator's voice-to-text posts. The distribution shows substantial heterogeneity, with most creators leaning personal while a smaller subset consistently posts highly promotional content.

OLS Regression

Statistically significant negative relationship: mean promotional probability is associated with lower creator median like rates (β = −0.018, p < 0.001), after adjusting for creator activity and text length. In contrast, creators with more posts (β = 0.0031, p < 0.001) and longer typical text (β = 9.96e−06, p = 0.001) tend to have slightly higher median engagement.

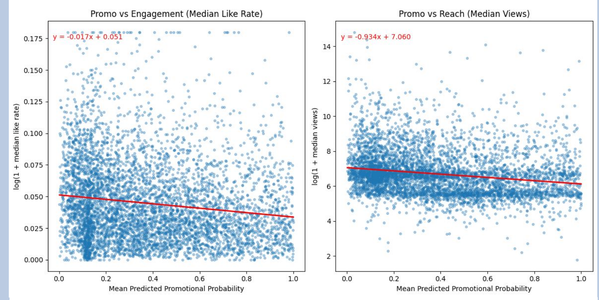

Figure 5. Higher Promotional Emphasis Is Associated with Lower Creator Engagement and Reach. Each point represents one mompreneur creator, plotted by mean predicted promotional probability (x-axis) and median engagement outcomes (y-axis, logged). **Left:** median like rate. **Right:** median view count. In both panels, the fitted line indicates a negative association: creators whose content is more promotional on average tend to receive fewer likes and views, suggesting reduced audience interaction and reach as promotional emphasis increases.

Conclusions

1. Mompreneur storytelling in the #SAHM community is **rarely purely personal or purely promotional**
   a. Most posts fall along a **continuum**, blending both styles
   b. Creators adjust this balance **over time**
2. **Higher promotional** emphasis is associated with significantly **lower engagement**, adjusting for posting volume, text length
   a. Creators who maintain primarily personal style, while selectively incorporating promotion can achieve more stable engagement
3. **Methodological contribution**
   a. Using TikTok's auto-generated voice-to-text captures spoken storytelling
   b. Hashtags are often optimized for algorithmic reach, while voice-to-text reflects richer narrative intent

Limitations

- Voice-to-text coverage represents only a subset of all #SAHM posts (~10%), and automatic transcription may introduce noise
- Engagement rates do not capture sentiment, long-term follower growth, or off-platform outcomes.

Future Direction

- Investigate how TikTok's recommendation algorithm differentially amplifies personal versus promotional storytelling styles.
- Incorporate richer linguistic features (sentiment, persuasion markers) to further differentiate mixed storytelling styles.
- Conduct longitudinal creator-level analyses linking storytelling shifts to audience growth and retention.

Acknowledgements

Thanks to Professor Eni Mustafaraj, and students in the fall 2025 DS capstone course.

References

1. Abidin, C. (2016). "Aren't These Just Young, Rich Women Doing Vain Things Online?": Influencer Selfies as Subversive Frivolity. Social Media + Society, 2(2). https://doi.org/10.1177/2056305116641342 (Original work published 2016)