Comparative Analysis:

Threads and Twitter Reviews

Prepared for: Prof. Ning Rui

Prepared by: Sai Rachana Kandikattu,

Snehitha Tadapaneni,

Haeyeon Jeong

December 05, 2025

I. Introduction

Social media platforms are central to shaping public discourse, making the analysis of user sentiment critical for understanding platform success and user satisfaction. The period of July 2023 was marked by simultaneous, high-impact events: the launch of Meta's Threads on July 5 , Twitter's rebranding to X , and the introduction of controversial post-viewing limits. This confluence of events created a unique opportunity to conduct a comparative analysis of user perception toward the new competitor (Threads) and the incumbent platform undergoing radical changes (Twitter).

The research objective is to understand how users perceived both platforms during this critical period using a rigorous Natural Language Processing (NLP) pipeline. We address this by performing comparative sentiment and topic modeling analyses. Our study utilizes a joint methodological approach, evaluating both classical NLP techniques (TF-IDF + Logistic Regression, LDA, NMF) and transformer-based models (Fine-tuned DistilBERT, BERTopic). This comparison allows us to assess model efficacy on noisy, short-form social media data while simultaneously extracting core insights into user experience.

To guide this comparative analysis, we formulate the following research questions:

- RQ1: How does user sentiment differ between Threads and Twitter?
- RQ2: How reliable are VADER pseudo-labels vs human annotations?
- RQ3: How do TF-IDF + Logistic Regression and DistilBERT perform?
- RQ4: How do LDA, NMF, and BERTopic differ in topic coherence and insights?

This work offers a systematic comparison of sentiment patterns and thematic content, providing actionable insights into early platform adoption and user experience dynamics.

II. Data

Two publicly available Kaggle datasets were selected for analysis, ensuring temporal alignment and comparability.

Threads Dataset:

- Source: Kaggle
- Total reviews: 32,910
- Period: July 2023 (launch month)
- Features:
  - review_description (text)
  - rating
  - review_date
  - source (Google Play or App Store)

Twitter (X) Dataset

- Source: Kaggle
- Total reviews: 34,788
- Period: July 2023
- Features:
  - review_text
  - review_rating
  - review_timestamp
  - review_likes, author_app_version

Both datasets contain short, informal, multilingual reviews with noise such as emojis, repeated content, typos, and inconsistent structure, typical of app-store feedback.

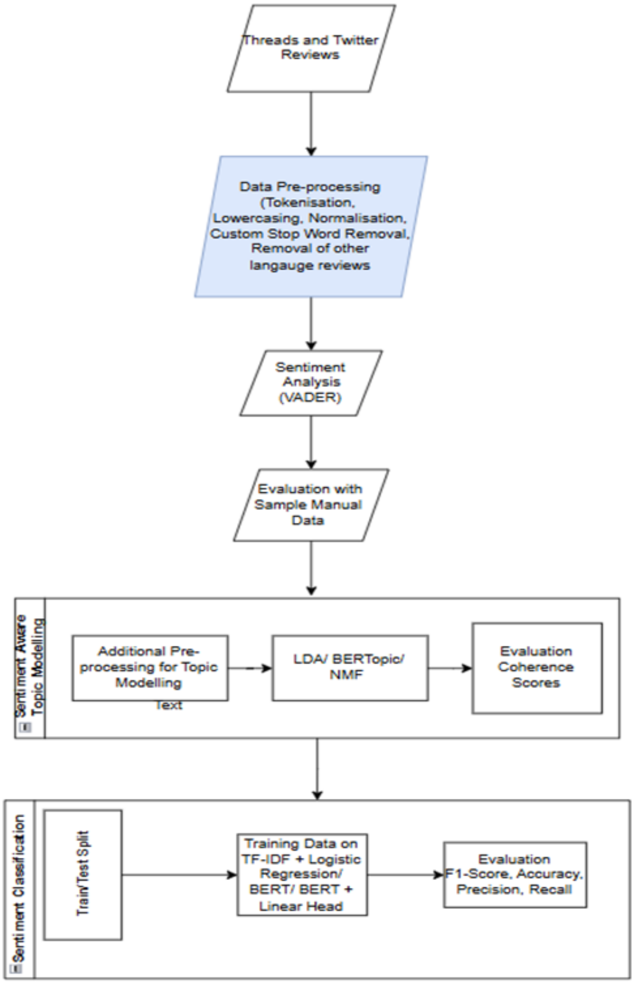
III. Methodology and Experimental Setup

The analysis was conducted using a joint pipeline of classical and deep learning models.

A. NLP Pipeline Overview

The pipeline processes data through sequential steps of labeling, classification, and topic modeling (Figure 1).

Figure 1
NLP Model Pipeline


Threads and Twitter Reviews
Data Pre-processing (Tokenisation, Lowercasing, Normalisation, Custom Stop Word Removal, Removal of other langauge reviews
Sentiment Analysis (VADER)
Evaluation with Sample Manual Data
Sentiment Aware Topic Modelling
Additional Pre-processing for Topic Modelling
Text
LDA/ BERTopic/ NMF
Evaluation Coherence Scores
Sentiment Classification
Train/Test Split
Training Data on TF-IDF + Logistic Regression/ BERT/ BERT + Linear Head
Evaluation F1-Score, Accuracy, Precision, Recall

B. Data Cleaning Pipeline

A unified preprocessing pipeline was executed identically on both datasets:

1. Duplicate Removal: Removed raw text duplicates and long-review duplicates (>15 words).

2. Brand Normalization: Standardized platform names (e.g., "twt," "x" → "twitter"; "treads" → "threads").

3. Text Normalization: Lowercased text, removed URLs, emojis, special characters, and normalized whitespace.

4. Stop-words & Lemmatization: Removed English stopwords but retained negations ("not," "never"), and lemmatized words

5. Minimal Language Filtering: Used langdetect to remove frequently misdetected languages, keeping English and "unknown" content.

The final dataset sizes were 29,646 for Threads and 29,611 for Twitter.

C. Sentiment Labeling (VADER + KNN)

VADER (Valence Aware Dictionary and Sentiment Reasoner) was used, as a lexicon-and-rule-based sentiment analyzer, to compute a normalized compound score $c \in [-1, +1]$.

We performed a reliability check on approximately 3,000 manually labeled samples to assess VADER's alignment with human judgment.

While standard thresholds use +0.5 and −0.5, we refined these using KNN distance patterns to suit this dataset. The optimal sentiment cut-off thresholds were determined using KNN distribution analysis for pseudo-label generation:

- Positive: c ≥ +0.25
- Neutral: between -0.25 and +0.25
- Negative: c ≤ 0.25

This adjustment improved separation between classes for short, noisy reviews.

D. Sentiment-Aware Topic Modeling

Prior to topic modeling, additional preprocessing was performed, including Bigram detection and selecting the representation (Bag-of-Words, TF-IDF, or Embeddings). This enhances topic coherence by treating multi-word concepts as single tokens.

- Bigram Modeling Example: To capture multi-word concepts, we construct bigrams using the Gensim Phrases model (min_count=10, threshold=10) and convert with a Phraser model. This transforms patterns such as:
    - “social media” → “social_media”
    - “user interface” → “user_interface”
    - “new update” → “new_update”

We compared LDA, NMF, and BERTopic.

- LDA (Latent Dirichlet Allocation): Used as a probabilistic, bag-of-words baseline.

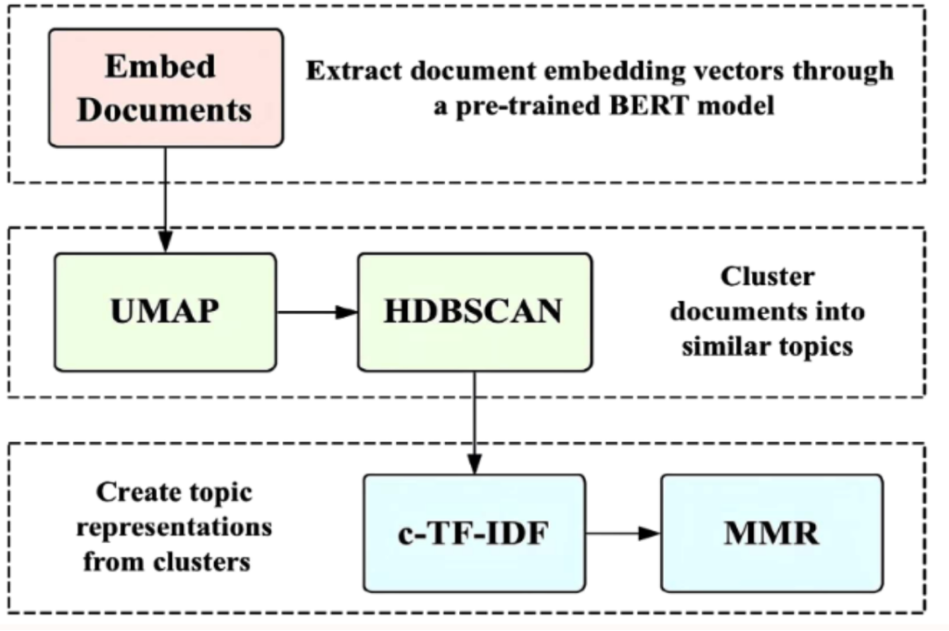
- NMF (Non-negative Matrix Factorization): Used TF-IDF matrix factorization, making it effective for sparse texts.
- BERTopic (Modern Embedding-Based Approach): Selected for its advanced architecture combining Sentence-BERT, UMAP, HDBSCAN, and c-TF-IDF

BERTopic Architecture and Superior Performance

BERTopic was selected as the Best Topic Model Architecture due to its modular design and superior coherence score on both datasets. This performance edge is attributed to its ability to leverage contextual embeddings for short, noisy social media text (Xenoss, 2025).

The BERTopic Architecture follows this embedding-based pipeline (Figure 2):

Figure 2
BERTopic Architecture


Embed Documents
Extract document embedding vectors through a pre-trained BERT model
UMAP
HDBSCAN
Cluster documents into similar topics
Create topic representations from clusters
c-TF-IDF
MMR

Note. This figure illustrates the modular three-step BERTopic pipeline: Embedding, Clustering, and Topic Representation. Adapted from BERTopic Documentation (n.d.).

1. Embed Documents: Extracts document embedding vectors through a pre-trained BERT model. This provides a semantic representation of each review.

2. UMAP (Dimensionality Reduction): Reduces vector size into a smaller space while preserving structure.

3. HDBSCAN (Clustering): Performs density-based clustering to find dense groups of similar reviews.

4. c-TF-IDF (Topic Extraction): Identifies the most representative words for each cluster.

E. Sentiment Classification Models

The following models were implemented and trained using the VADER-generated pseudo-labels:

1. TF-IDF + Logistic Regression (Baseline): This traditional model serves as the performance baseline.

2. Fine-tuned DistilBERT (Advanced): This transformer model was chosen for its ability to capture semantic and contextual information. DistilBERT is a smaller, faster, and lighter version of the BERT model, which makes it computationally efficient for classification.

- Technical Background (Transformer Architecture): DistilBERT utilizes the Transformer architecture and the self-attention mechanism to understand the context of every word in relation to all other words in the sentence. The foundation of this architecture is the multi-head attention mechanism. The output of the scaled dot-product attention for a query Q, key K, and value V is given by the standard Transformer equation (Vaswani et al., 2017):

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

where d_k is the dimension of the key vectors. This capability allows the model to capture complex semantic and syntactic relationships necessary for interpreting short, informal social media text.

- Hyperparameters (Initial Configuration):

  Threads:

  - num_train_epochs=3
  - learning_rate=2e-5
  - per_device_train_batch_size=16
  - per_device_eval_batch_size=16
  - weight_decay=0.01

  Twitter:

  - num_train_epochs=3
  - learning_rate=1e-5
  - per_device_train_batch_size=16
  - per_device_eval_batch_size=16
  - weight_decay=0.01

- Hyperparameter Tuning: The DistilBERT model was optimized using Hugging Face's Optuna tool, which searched through ≈20 combinations of parameters. The hyperparameters tuned included: Learning rate, Number of epochs, Batch size, Weight decay, and Warmup steps.

F. Performance Evaluation

The performance of all models was judged using the following metrics:

- Sentiment Classification: Model reliability was measured using Accuracy and the Weighted F1-Score.
- Topic Modeling: Topic quality and interpretability were measured using the Coherence Score.

IV. Results and Discussion

A. Sentiment Comparison Insights (RQ1)

Table 3
Twitter and Threads Sentiment Comparison

Sentiment	Threads	Twitter
Positive	16K	11K
Negative	9K	10K
Neutral	4K	9K

As shown in Table 3, the sentiment distribution based on VADER pseudo-labels showed that Threads generated a substantially higher number of positive reviews (≈ 16k) compared to Twitter (≈ 11k). Neutral reviews were markedly more prevalent on Twitter (≈ 9k) than on Threads (≈ 4k). Negative reviews were similar in magnitude (≈ 9k to ≈ 10k), indicating comparable dissatisfaction. A key finding was that many reviews which were complaints about app glitches and bugs on Threads were classified as neutral by VADER due to the absence of explicit negative words.

B. Sentiment Classification Model Performance (RQ2 & RQ3)

The reliability check against manual annotations (≈ 3,000 samples) showed that VADER demonstrated moderate agreement across both platforms, though performance notably differed. This evaluation reflected a practical estimate of VADER's alignment with human judgment rather than an absolute ground truth.

On Threads, VADER demonstrated moderate reliability, with model performance showing high recall for positive sentiment (0.93) but substantially weaker recall for negative (0.40) and neutral (0.39) classes (Table 4).

Table 4
Threads VADER Reliability against Manual Labels

Metric	Positive	Negative	Neutral
Accuracy (Overall)	0.54	0.54	0.54
F1-score	0.65	0.45	0.47

For Twitter, performance was comparatively stronger, achieving an overall accuracy of 0.60 (Table 5). These trends imply that VADER aligns more effectively with human judgments on Twitter, likely due to its original calibration on short, informal, and emoji-rich text characteristic of that platform.

Table 5
Twitter VADER Reliability against Manual Labels

Metric	Positive	Negative	Neutral
Accuracy (Overall)	0.60	0.60	0.60

F1-score	0.76	0.65	0.19

The models' classification performance is summarized in Table 6.

Table 6
Setiment Classfication Performance

Model	Threads Accuracy	Threads Weighted F1	Twitter Accuracy	Twitter Weighted F1
TF–IDF + Logistic Regression	0.74	0.69	0.84	0.83
DistilBERT	0.86	0.86	0.93	0.92

- DistilBERT significantly outperformed Logistic Regression on both platforms.
- DistilBERT's final performance was higher on Twitter (Acc 0.93, F1 0.92) than Threads (Acc 0.86, F1 0.82), primarily due to the less noisy VADER labels on the Twitter dataset.
- Finding optimal parameters and fine-tuning improved the model performance for both Threads and Twitter.

C. Topic Modeling Comparison and Insights (RQ4)

As shown in Table 7, BERTopic yielded the highest coherence scores on both platforms: Threads (≈ 0.53) and Twitter (≈ 0.52). This compares favorably to LDA (Threads ≈ 0.52, Twitter ≈ 0.50) and NMF (Threads ≈ 0.46, Twitter ≈ 0.43). , making it the most suitable model for short, noisy social media text.

Table 7
Topic Model Coherence Comparison

Model	Threads (Avg. Coherence Score)	Twitter (Avg. Coherence Score)
LDA	≈ 0.52	≈ 0.50
NMF	≈ 0.46	≈0.43

BERTopic	≈ 0.53	≈ 0.52

1. Latent Dirichlet Allocation (LDA)

LDA provided a useful baseline, capturing broad themes with coherence values ranging from 0.43 to 0.57.

- Threads LDA Topics: Topics centered on general satisfaction, feature needs, and technical stability.
    - Positive (Best k=4): Focused on General positivity (e.g., *super*, *awesome*), Comparisons favouring Threads (*twitter*, *better*), Polite feature requests (*feature*, *need*), and Platform approval.
    - Negative (Best k=3): Clustered around clear pain points: Crashes and reliability failures (*worst*, *post*, *crash*), Low quality experience/missing features, and Strong rejection of the product (*useless*, *boring*).
- Twitter LDA Topics: Discussions were less about basic functionality and more about rebranding, leadership, and platform identity.
    - Positive (Best k=3): Included General enthusiasm, Positive comparisons/improvements (e.g., *free_speech*, *better*), and Satisfaction with changes.
    - Negative (Best k=3): Highly emotional and linked to recent changes: Usability/update complaints (*no*, *worse*, *limit*), Strong disapproval of rebranding (*suck*, *logo*, *bird*), and Leadership-focused criticism (*bad*, *elon*, *ruined*, *musk*).

LDA clearly distinguished the major drivers of sentiment: Threads negativity centers on crashes/missing features, while Twitter negativity focuses on rebranding/leadership decisions.

2. Non-Negative Matrix Factorization (NMF)

NMF, operating on TF-IDF weighting, produced clearer and more interpretable topics than LDA, especially suitable for sparse texts like reviews.

- Threads NMF Topics: NMF captured a wider variety of frustrations in negative sentiment ($k=9$).
    - Positive (Best k=6): Reflected General Impressions, Competitor Comparisons (*facebook*, *instagram*), and Functionality Mentions.
    - Negative (Best k=9): Topics were complex, including: App Not Working (*not_working*, *glitch*), UI & Privacy Concerns, Missing Functions, Strong Dissatisfaction (*rubbish*, *useless*), and Technical Failures (*crash*, *upload*).
- Twitter NMF Topics: Topics were sharp, driven by unique terms.
    - Negative (Best k=3): Themes were: Update-Related Frustrations, Rebranding Disapproval (*suck*, *name*, *logo*), and Leadership Criticism (*bad*, *elon*, *ruined*, *musk*).

NMF reduced the overlap of generic terms seen in LDA and highlighted more discriminative terms, capturing more distinct sentiment-specific themes.

3. BERTopic Insights

BERTopic, the most accurate model, revealed the clearest differences:

- Threads (New-platform stability): Positive sentiment was simple (*good app*); negative sentiment was dominated by crashes, missing features, and bugs, indicating criticism tied to the platform's immaturity.

- Twitter (Identity & Leadership): Positive sentiment included ideological support for free speech; negative sentiment was intense, driven by Musk decisions, policy changes, rate limits, and rebranding anger.

The core finding is that Threads is criticized for what it lacks, while Twitter is criticized for what it has become.

Table 6

Key Topic Insights (BERTopic)

Platform	***Sentiment***	***Major Topics Extracted***
Threads *(New-platform stability)*	***Positive***	1. General positive impressions - *good, nice, best, awesome(These reviews were about how good or amazing the app is)* 2. Positive Twitter comparisons - *twitter, better, app (These reviews were about how threads is better than twitter)* 3. Positive Experience - *app, good, nice(these reviews were about how the app is good or nice)* 4. Copy/clone comments (positive or sarcastic) - *copy, copycat, original(These reviews were how the app is a good copy or a copy is just a copy so these were either positive or sarcastic which could be hard to identify with words such as nice, good)*
	Neutral	1. Platform comparisons or suggestions - *twitter, instagram, account (these reviews were about how threads and instagram are linked requesting for updates or suggestions on how it could be improved)* 2. Functionality & bug mentions - *app, crashing, glitching, bug (These reviews were mostly about users explaining glistches and bugs with the app)* 3. Leadership mentions - *elon, musk, zuck (These reviews were mostly just mentioning elon musk or mark zuckerberg)*

		4. Copying discussions - *copy, copied, clone(these reviews were how threads is just a clone or copy of twitter)*
	Negative	1. Lacking aspects - *app, account, not(These reviews were users complaining about how the threads is still not good enough and lacking in various aspects)* 2. Bad experience - *bad, boring, useless(These reviews were about users expressing disappointment with how threads was boring, useless etc)* 3. Leadership/platform direction critique - *elon, musk, zuck, hate (These reviews were just criticizing different leaders)* 4. Accusing Threads of cheating- *cheating,competition, not ,fine(These reviews were how threads is not a good competitor as its just cheating)*
Twitter *(Identity & Leadership)*	***Positive***	1. How Amazing twitter is compared to X - *app, good, best(These reviews were mostly about how good or amazing the twitter app was but how people feel weird but about changes made to make it into X)* 2. Simple praise - *good, nice, great, excellent(These reviews just explained how good the app or amazing the app is)* 3. Free speech appreciation of X - *speech, free, freedom, elon (These reviews were specifically about X after it has been changed and the appreciation of free speech where people could share their unfiltered opinions)* 4. Account help requests - *account, help, please, number(These reviews were mostly about requesting help for their accounts due to them being hacked, suspended etc) These reviews were filtered as positive by VADER as they were not necessarily negative or neutral*
	Neutral	1. Rebranding/name-change remarks - *twitter, app, (these reviews are just about how twitter changed)* 2. App quality over time - *trash, used, better(These reviews are mostly about how the ads, bugs and*

		how app degraded) -These seem to be bit on the negative sentiment) 3. Elon mentions (neutral tone) - *elon, musk (These are reviews which are just mentioning elon musk)* 4. Twitter Logo change - *bird, blue, back (These reviews were mostly about asking the bird logo to be back which could also be considered neutral sentiment as there is not much disappointment expressed)*
	Negative	1. Platform decline due to updates - account*, app, updates (These reviews outright express the disappointment towards the app updates and their accounts being suspended)* 2. Disappointment with Elon - *elon, ruined, sucks, worst (These are reviews were specifically targeted towards elon and how he ruined the app)* 3. Bad experiences - *bad, worse, sucks, worst (These are reviews about how the app is not runied and how users are not at all happy with the change)* 4. Strong ideological/political frustration - *racism, hate, biased(These reviews were about how there is more hatred and racism on the platform)*

V. Conclusion and Recommendations

A. Conclusion

The analysis showed that Threads faced criticism focused on functional immaturity (stability, missing features). Twitter's sentiment was highly polarized, with criticism intensely focused on leadership and policy direction (Musk changes, rebranding). Technically, DistilBERT performed better by capturing context and emotion , and BERTopic produced the clearest topics for short, noisy reviews.

B. Limitations and Future Work

- Limitations: VADER struggles with mixed or nuanced sentiment. Pseudo-labels introduced noise into model training.
- Future Work: Explore alternative topic extraction models such as Top2Vec and investigate multi-label classification for reviews expressing multiple themes.

C. Recommendations (Business Insights)

- Threads (New Products): New products must prioritize delivering a stable core experience and essential functionality first.
- Twitter/X (Major Updates): Major updates or rebranding efforts require careful, transparent communication to maintain user trust.

VI. Appendix

Appendix A: Topic Model Performance

Table 7

Coherence Scores Comparision: Threads

Scores Comparision	**Various Topic Modelling Techniques**		
	Model	***Coherence Scores***	***Topic Distribution***

	LDA	Pos = 0.56
		Neu = 0.58
		Neg = 0.52
		(Avg = 0.58)
			Best for understanding broad themes, less effective for short text

	NMF	Pos = 0.47
		Neu = 0.45
		Neg = 0.48
		(Avg = 0.46)
			Performs strongly on neutral and positive, weaker on negative (short text)

	BERTopic	Pos = 0.52
		Neu = 0.59
		Neg = 0.49
		(Avg = 0.53)
			Automatically finds topic structure; best for fine-grained distinctions

Table 8

Coherence Scores Comparision: Twitter

Scores Comparision	**Various Topic Modelling Techniques**		
	Model	***Coherence Scores***	***Topic Distribution***

	LDA	Pos = 0.54
		Neu = 0.57
		Neg = 0.40
		(Avg = 0.50)
			Produces broad, interpretable topics but sometimes includes generic filler words

	NMF	Pos = 0.45
		Neu = 0.45
		Neg = 0.40
		(Avg = 0.43)
			Generates sharper topics driven by TF–IDF weighting

	BERTopic	Pos = 0.60
		Neu = 0.55
		Neg = 0.43
		(Avg = 0.52)
			Produces the most semantically coherent and fine-grained topics

Appendix B: LDA and NMF Topic Summaries

Table 9

LDA Topic Summaries

Platform	***Sentiment***	***Key Topic Content Summary***
Threads	***Positive*** **(best k=4)**	• *General positivity and praise* – terms such as *super, awesome, good, best, amazing, excellent* show broad satisfaction and excitement. • *Comparisons favouring Threads* – words like *twitter, better, nice, great, love* indicate that many users explicitly compare Threads to Twitter and often prefer Threads.

		• *Polite feature requests – feature, need, option, add, feed, follow* show that even satisfied users suggest improvements to feed sorting and following options. • *Platform approval* – combinations of *cool, ok, pretty, social_medium, platform* reinforce that early adopters see Threads as a promising app.
	***Neutral* (best k=3)**	• *Platform comparisons and copying discourse – twitter, copy, facebook, instagram, first* reflect discussion of Threads as a "copy" of Twitter and general commentary on Meta's strategy. • *Functional/technical notes – need, post, use, bug, glitch, see, even, work* capture observations about bugs, glitches and missing controls. • *Twitter vs Elon Musk narrative – twitter, elon_musk, copy_cat, clone, buggy* revolve around discussions of Musk, Twitter and the idea that Threads cloned Twitter.
	***Negative* (best k=3)**	• *Crashes and reliability failures – worst, post, crash, error, problem, anything, work, time, bug, glitch* highlight instability and failed posting. • *Low quality experience and missing features – bad, twitter, instagram, feed, use, people, need, delete, want* mix dissatisfaction with missing functionality and unfavourable comparisons. • *Strong rejection of the product – useless, boring, poor, not_good, waste, fake, disgusting, awful, ugly* reflect users who see Threads as a low-quality or unnecessary alternative.
Twitter	***Positive*** **(best k=3)**	• *General enthusiasm & enjoyment - love, excellent, best_social, video, keep, lmao*; users praise the content, interactions, and overall experience. • *Positive comparisons & improvements - better, awesome, free_speech, amazing, cool, perfect*; reviews highlight perceived improvements and features (especially around free speech) compared to previous versions or competitors. • *Satisfaction with functionality or changes - good, great, like, elon, name, change, logo*; some users

		explicitly approve of changes introduced under Elon Musk, including logo/name changes.
	Neutral **(best k=6)**	• *Leadership & feature commentary* - *elon_musk, cringe, deleted, feature, uninstalled*; users describe leadership decisions or removed features without clear sentiment. • *Platform behaviour and quality* - *elon, use, ruined, platform, anymore*; descriptive remarks on how the platform now behaves. • *Functional notes on usage* - *tweet, trash, video, limit, see, update*; matter-of-fact comments on rate limits, tweet visibility and video behaviour. • *Rebranding & logo discussion* - words about the bird logo, bringing it back, and rebranding; users notice the change but do not always attach strong emotion. • *Updates & usability* - *update, top, phone, everything, new*; observations about how recent updates affect day-to-day use. • *Logo/name change mentions* - *change, name, old, logo, bird, miss*; neutral nostalgia for the previous brand identity.
	Negative **(best k=3)**	• *Usability & update complaints* - *no, worse, not, update, limit, post, people*; users report that updates and rate limits make tweeting harder or less enjoyable. • *Strong disapproval of rebranding* - *suck, name, terrible, logo, bird, ugly, back*; many reviews strongly dislike the new “X” name and logo and want the old bird branding back. • *Leadership-focused criticism* - *bad, elon, worst, ruined, musk, awful*; users attribute perceived decline of the platform directly to leadership decisions.

Table 10

NMF Topic Summaries

Platform	***Sentiment***	***Key Topic Content Summary***
Threads	***Positive*** (Best k = 6)	• *General Impressions* - new, use, feature, need • *Interface & Platform Notes* - see, ui, platform • *Competitor Comparisons* - facebook, instagram, alternative • *Meta Ecosystem Context* - social, platform, world • *Functionality Mentions* - work, job, start, feature • *Light Positive Reactions* - cool, wow, amazing, love
	Neutral (Best k = 6)	• *Copying/Clone Commentary* - cheap, clone, copied, twitter • *Account & Login Activities* - login, delete, create, sign • *Competitor Mentions* - mark, zuck, tweeter • *Feed & Usage Observations* - feed, see, post, work • *Minor Technical Issues* - glitching, bug, not_working • *Routine App Interactions* - write, review, comment, time
	Negative (Best k = 9)	• *App Not Working* - not_working, glitch, install, ui • *UI & Privacy Concerns* - screen, privacy, content • *Missing Functions* - post, see, need, delete • *Strong Dissatisfaction* - rubbish, pathetic, useless • *Copying/Clone Complaints* - copy, copying, cheap, copy_twitter • *Design Issues* - boring, nothing_new, poor • *Feed/Discovery Problems* - feed, trending, hashtags • *Quality & Competitor Comparison* - clone, fake, twitter • *Technical Failures* - crash, upload, picture, try
Twitter	***Positive*** `(k ≈ 3–5)`	• *General Enjoyment* — love, excellent, best_social • *Feature Appreciation / Improvements* — better, awesome, free_speech • *Satisfaction with Updates* — good, great, name, change, logo
	Neutral `(k ≈ 6)`	• *Leadership & Feature Mentions* — elon_musk, feature • *Platform Behavior* — ruined, platform, anymore • *Tweet/Video/Limit Notes* — tweet, limit, video • *Rebranding Discussions* — bird, back, rebranding

		- *Update Notes* — update, phone - *Logo/Name Change* — change, name, logo
	Negative (k ≈ 3)	- *Update-Related Frustrations* — no, worse, limit, post - *Rebranding Disapproval* — suck, name, logo, bird - *Leadership Criticism* — bad, elon, ruined, musk

Appendix C: VADER Pseudo-Label Reliability (Manual Labelling)

To assess the reliability of the VADER-generated pseudo-labels, we compared them against a manually annotated subset of approximately 3,000 samples. Although human labeling is inherently subjective, the evaluation reflects a practical estimate of VADER's alignment with human judgment

Table 11

Threads VADER Reliability against Manual Labels

Metric	***Positive***	***Negative***	***Neutral***
Accuracy (Overall)	*0.54*	*0.54*	*0.54*
F1-score	*0.65*	*0.45*	*0.47*

On Threads, VADER demonstrated moderate reliability, with model performance showing high recall for positive sentiment (0.93) but substantially weaker recall for negative (0.40) and neutral (0.39) classes.

Table 12

Twitter VADER Reliability against Manual Labels

Metric	***Positive***	***Negative***	***Neutral***
Accuracy (Overall)	*0.60*	*0.60*	*0.60*
F1-score	*0.76*	*0.65*	*0.19*

For Twitter, performance was comparatively stronger, achieving an overall accuracy of 0.60. These trends imply that VADER aligns more effectively with human judgments on Twitter, likely due to its original calibration on short, informal, and emoji-rich text characteristic of that platform.

References

BERTopic. (n.d.). *The BERTopic Algorithm*. BERTopic Documentation. Retrieved from https://bertopic.com/

Grootendorst, M. (2022). **BERTopic: Neural topic modeling with a class-based TF-IDF procedure**. *arXiv preprint arXiv:2203.05794*. *https://arxiv.org/abs/2203.05794*

Kaggle. (2023). *Threads: An Instagram App Reviews* [Data set]. Kaggle. https://www.kaggle.com/datasets/saloni1712/threads-an-instagram-app-reviews

Kaggle. (2023). *2 Million X (Twitter) Google Reviews* [Data set]. Kaggle. https://www.kaggle.com/datasets/bwandowando/2-million-formerly-twitter-google-reviews

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). **Attention is all you need**. In *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Xenoss. (2025, April 7). **Topic modeling techniques: LDA, NMF, Top2Vec & BERTopic**. Xenoss Blog. https://xenoss.io/blog/topic-modeling-techniques-comparison