SOCIAL MEDIA TRENDS ANALYSIS: DATA-DRIVEN INVESTIGATION OF ENGAGEMENT PATTERNS USING MACHINE LEARNING

Mirco Fernando

BSc (Hons) – Software Engineering (Reading)

Zuu Crew Academy

Module: Machine Learning Foundations

Abstract- **This project examined content trends across TikTok, Instagram, Twitter, and YouTube to determine the key factors for viral success and accurately predict engagement. Using a cleaned and feature-engineered multi-platform dataset, the study applied extensive exploratory analysis, supervised learning, and unsupervised clustering to uncover platform-specific behavior and engagement patterns. The results indicated that standard prediction models (regression) were not effective because social media data is complex and changes rapidly non-linear. In contrast, classification models worked Among them, XGBoost achieved the best regression performance overall, while CatBoost handled class imbalance most effectively in classification—producing more stable predictions across low, medium, and high engagement levels.. For strategic insight Unsupervised Learning, Dimensionality reduction and KMeans clustering successfully identified three clear audience segments, and DBSCAN located unique, unexpected patterns associated with viral content. The analysis proves that advanced grouping and classification methods are essential for understanding success across different social media platforms. The findings lead to practical, data-backed recommendations that marketers and creators can use to optimize their content strategy.**

Key Methodology

- Comprehensive feature engineering created 60+ predictive features
- Supervised learning: 4 regression tasks (Views/Likes/Shares/Comments) and engagement classification
- Unsupervised learning: PCA, t-SNE, UMAP for pattern discovery
- KMeans & DBSCAN for segmentation
- Model Evaluation metrics

I. INTRODUCTION

A. *Background*

Social media platforms such as YouTube, Instagram, Twitter, and TikTok have emerged as a key factor in the involvement of society. They now serve as the entry point for a variety of activities, including communication, entertainment, culture, and commerce (Paljug, 2025) As of 2025, five billion people worldwide—or 68% of the world's population—use social media, which is more than half of the world's population. engagement metrics such as views, likes, comments, and shares have become central indicators of content performance and audience behaviour. These metrics influence platform algorithms, shape visibility, and affect the success of creators and brands. Businesses' most powerful tool for marketing and branding is social media leveraging social media effectively can help businesses to thrive in growth building strong brand awareness and drive meaningful and more engagement with customers. (Joshi, 2024)

B. *Motivation*

In an increasingly competitive social media environment, businesses, creators, and marketers require reliable tools to predict engagement and identify patterns that influence content success. Manual analysis is insufficient due to the scale and complexity of the data. Machine learning offers the ability to model nonlinear behaviour, uncover hidden structures, and provide actionable insights that traditional methods cannot achieve. This motivates the need for a system that not only predicts engagement outcomes but also reveals underlying behavioural clusters, enabling data-driven decision-making for content optimization and trend analysis.

C. *Problem Statement*

Although social media platforms generate vast amounts of engagement data, deriving meaningful insights from it remains a challenge. Factors such as platform diversity, content type variation, regional differences, and user behaviour contribute to inconsistent engagement patterns. There is a lack of integrated analytical solutions capable of simultaneously:

- Predicting numerical engagement metrics,
- Classifying overall engagement performance, and
- Discovering hidden behavioural groups using unsupervised learning.

This project addresses the problem by developing a machine-learning framework capable of analysing, predicting, and interpreting social media engagement data.

II. DATA INVESTIGATION

A. Overview

The dataset used in this study contains a collection of social media posts sourced from multiple platforms, including YouTube, Instagram, Twitter, and TikTok. Each record represents a single content item along with its associated engagement statistics and contextual metadata. The dataset encompasses both categorical and numerical variables, enabling a rich analysis of platform behaviour, audience interaction, and content performance. This dataset consists of 10 feature columns. But when investigating thoroughly, this dataset consisted of many anomalies and needed feature engineering for model training lets discuss each.

- Categorical features – Platform, Content Type, Hashtag, Region.
- Numeric Features – Views, Likes, Shares, Comments
- Classification Target – Engagement Level

B. EDA (Exploratory Data Analysis)

Missing value analysis- Initial inspection revealed no missing data, removing the need for imputation. This improved the consistency of preprocessing and model training.

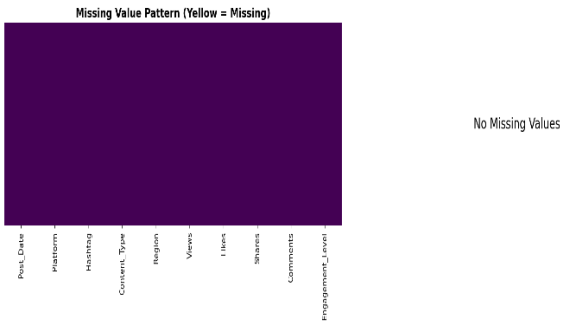


Figure –1. Missing value analysis

Anomaly investigation-The most critical part of a model development pipeline is the data preparation. I conducted a thorough investigate into the dataset and domain specific outliers. particularly in domains where user interactions follow strict behavioural rules. In social media, engagement occurs in a fixed sequence: a post must be **viewed** before it can be liked, shared, or commented on. Therefore, patterns such as **Likes exceeding Views** or Shares and Comments being higher than Views represent **domain-impossible values**. These anomalies likely created from data errors, artificial activity or data collection errors, and must be identified to ensure the dataset reflects realistic user behaviour.

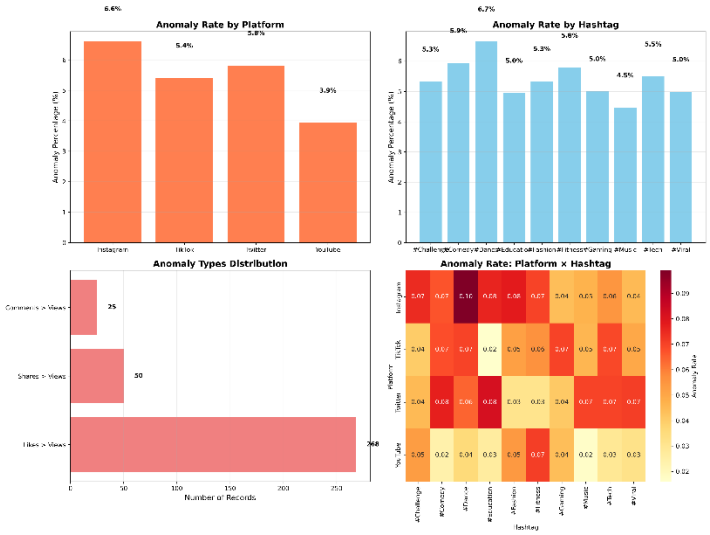


Figure 2 – Anomaly Investigation

This figure illustrates all the considered anomalies presented in all categories. The rows with anomalies are removed, Total – 270 were removed, And Content Type anomalies were also found where the content type does not match the Platform. There were total 1822 content mismatches, we can't remove them because it's a large amount of data. Instead, these mismatches were corrected by making assumptions, as an example if YouTube contain a content type "Reel" it is replaced to "Shorts". this will enhance model performance for real world data.

C. *Data visualization*

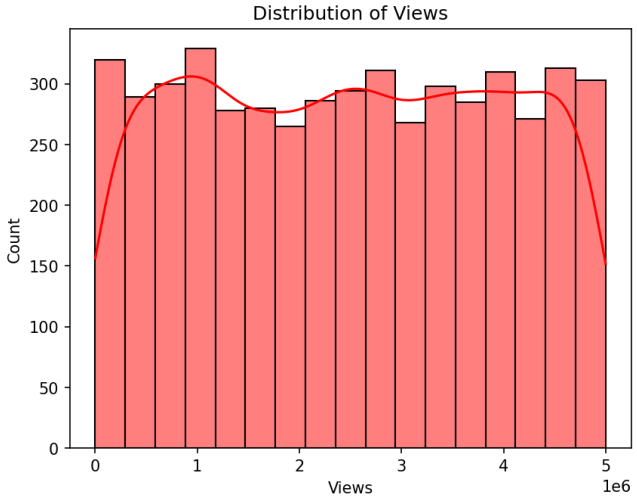


Figure 3 – Distribution of Views

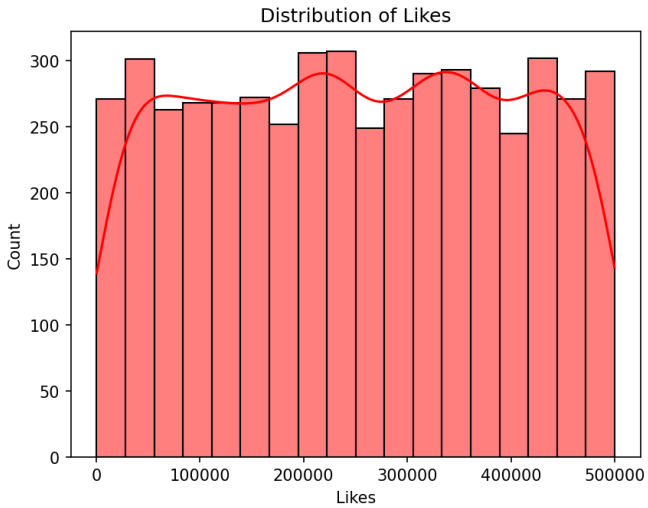


Figure 3 – Distribution of Likes

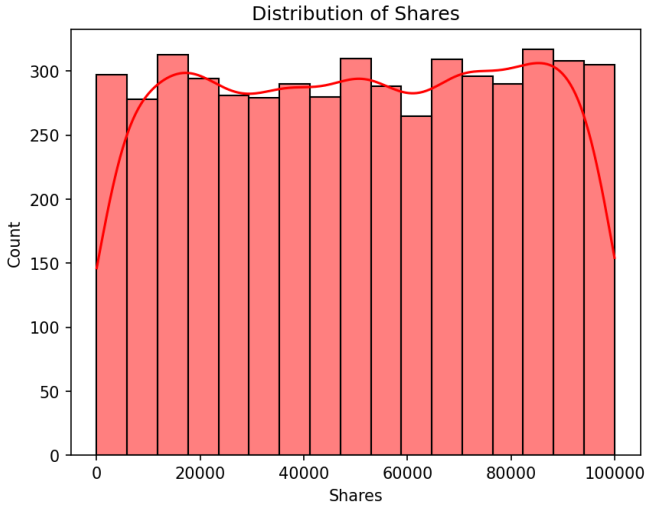


Figure 3 – Distribution of Shares

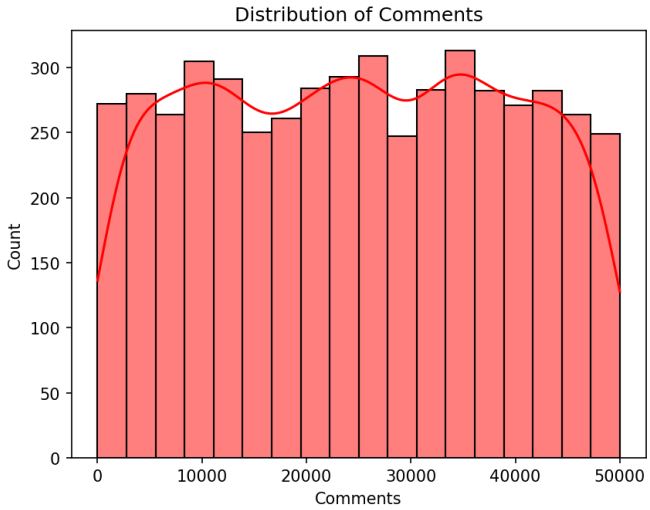


Figure 3 – Distribution of Comments

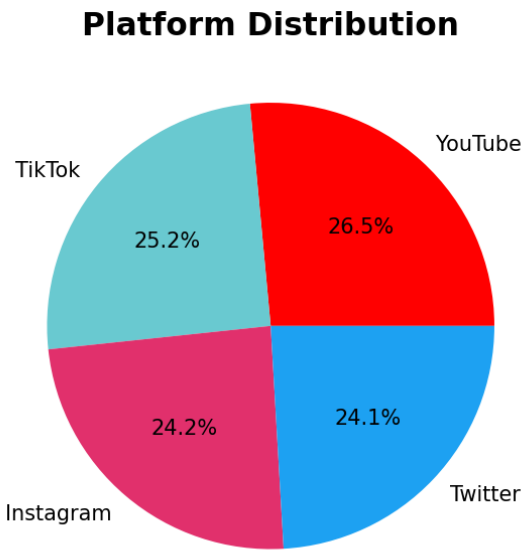


Figure 4 – Platform Distribution

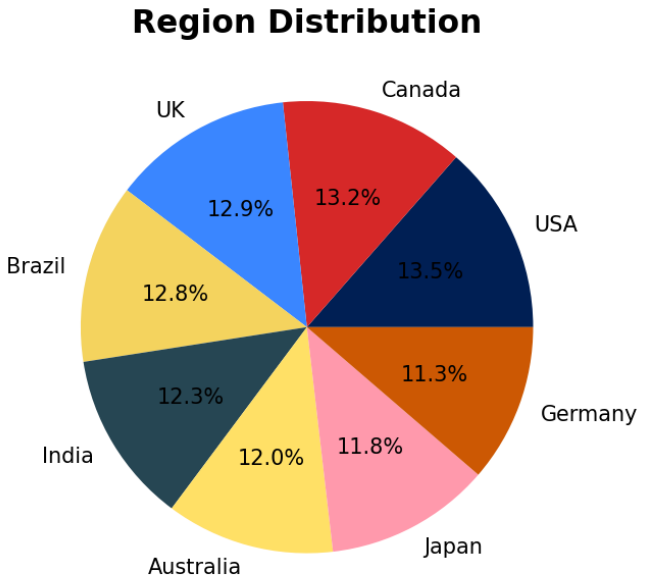


Figure 5 – Region Distribution

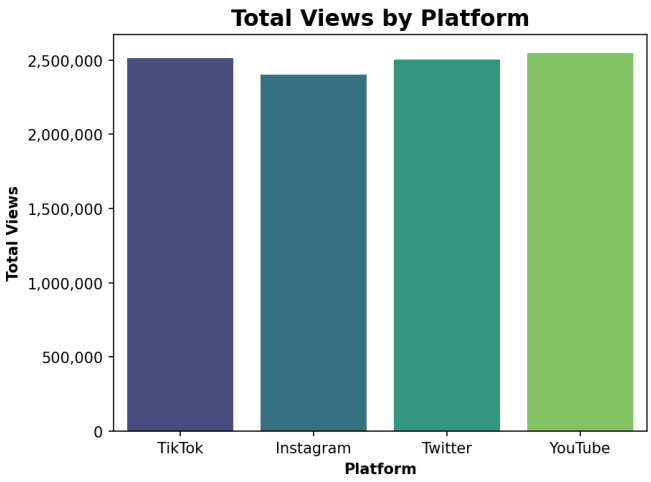
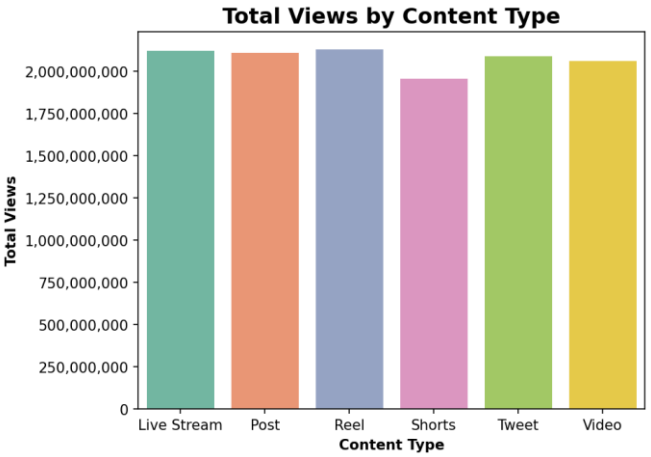
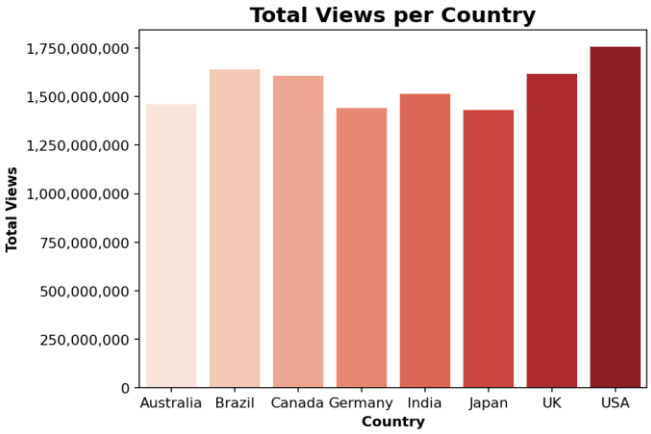
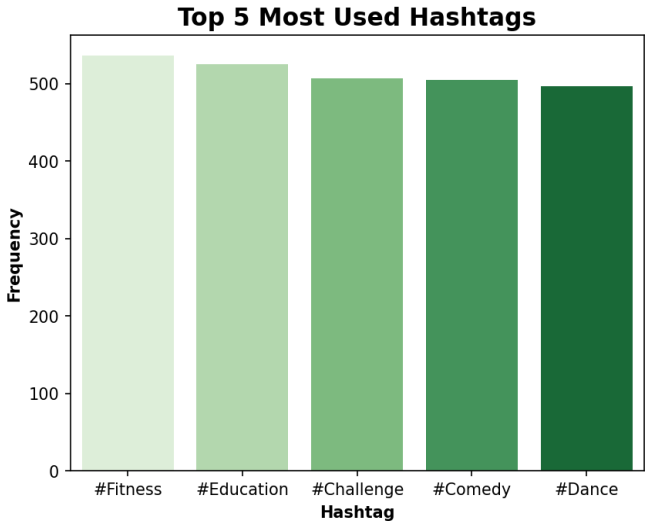


Figure 6 – Total views by platform







D. Feature engineering before SPLIT

To improve model performance and capture deeper engagement patterns, more than **seventy engineered features** were created. And more after splitting the data for model optimization. All features are available in the 'artifact's' directory. Among them a crucial feature is the Engagement Rate.

$$engagement_{rate} = \frac{Likes + Share + Comments}{Views}$$

This is important when analysing performance in digital marketing because it illustrates how your audience is interacting with your content and can be used to produce further strategies and insights. And it's used to calculate the correct Engagement Level.

Additionally, many other features are made for future interpretations and for Unsupervised Learning. These features will not be used in Supervised Learning because they will cause data leakage.

Data leakage- Data leakage is when your model learns from information it should not have access to, leading to overly optimistic and misleading performance.

The dataset includes an **Engagement Level** variable that is used as the classification target; however, this label is **not aligned with real-world platform-specific engagement standards**. The dataset calculates engagement levels using a **quantile-based method**, which distributes values evenly into "Low," "Medium," and "High" categories without

considering how engagement behaves on different platforms. Engagement expectations vary significantly across platforms due to differences in algorithms, user behaviour, and content formats, A newly calculated engagement level with these platform specific rates will be used as the classification target. Applying this will cause class imbalances due to increasing "High" Level this will be addressed later in the report.

Instagram

- Low: < 1% - engagement rate
- Medium: 1%–3%- engagement rate
- High: > 3%- engagement rate

TikTok

- Low: < 5%- engagement rate
- Medium: 5%–12%- engagement rate
- High: > 12%- engagement rate

YouTube

- Low: < 2%- engagement rate
- Medium: 2%–4%- engagement rate
- High: > 4%- engagement rate

Twitter/X

- Low: < 0.5%- engagement rate
- Medium: 0.5%–1.5%- engagement rate
- High: > 1.5%- engagement rate

III. METHODOLOGY

A. Data preprocessing

Data preprocessing is a critical stage in the machine learning pipeline, ensuring that the dataset is clean, consistent, and suitable for modelling. This project followed a structured and leakage-free preprocessing workflow consisting of outlier detection, feature engineering, encoding, scaling, feature selection, and artifact saving.

1. *Split Train/Test*

An **80/20 train-test split**, ensuring: Training data is used for fitting scalers, encoders, and model Test data is reserved strictly for evaluation. No data leakage occurs.

Unsupervised Learning Dataset- The full feature matrix (after preprocessing) is used for: PCA, t-SNE, UMAP, KMeans, DBSCAN

This enables better discovery of natural cluster structures across all posts.

- **Training set**: 3,784 samples
- **Test set**: 946 samples

Numeric Targets – Views, Shares, Comments, Shares

Classification Target – Engagement Level Cal(Derived from correct platform specific rates)

2. Outlier Detection

Outliers were evaluated using the **Interquartile Range (IQR) method**, applied only to training data to avoid leakage.

$$IQR = Q3 - Q1$$

$$Lower\ Bound = Q1 - 1.5 \times IQR$$

$$Upper\ Bound = Q3 + 1.5 \times IQR$$

No statistical outliers were detected in core engagement metrics. Engagement Rate values exceeding 100% were treated separately and capped at 100%, preventing distortion in regression and clustering.

3. Feature Engineering After Split

To enhance model performance and capture richer behavioural patterns, more than **seventy engineered features** were created. These were derived **after splitting** the data to avoid target leakage. But some features will be excluded due to high corelation between them.

1) *Average views on all ordinal columns*

Captures the typical popularity, helping the model understand which hashtags tend to get more views and other.

$$Avg_{Views}i = mean(Views\ for\ all\ posts\ with\ same\ Ordinal\ col(i)$$

2) *Frequency*

$$freq\ (i) = count\ of\ posts\ Ordinal\ col\ (i)$$

3) Date Since first post

$$days\_since\_first\_post(i) = Post\_Date - Earliest\ Post\ Date\ in\ Dataset$$

Some other features were made but they showed high corelation in the dataset.

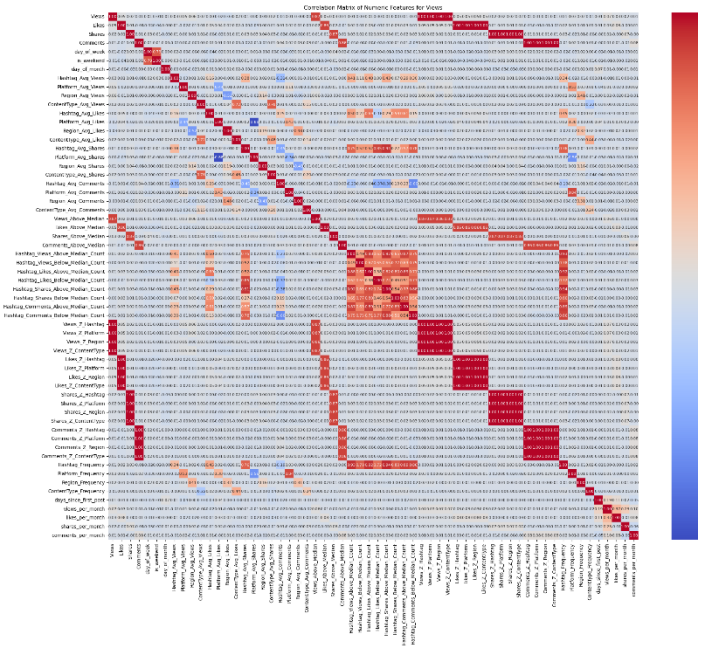
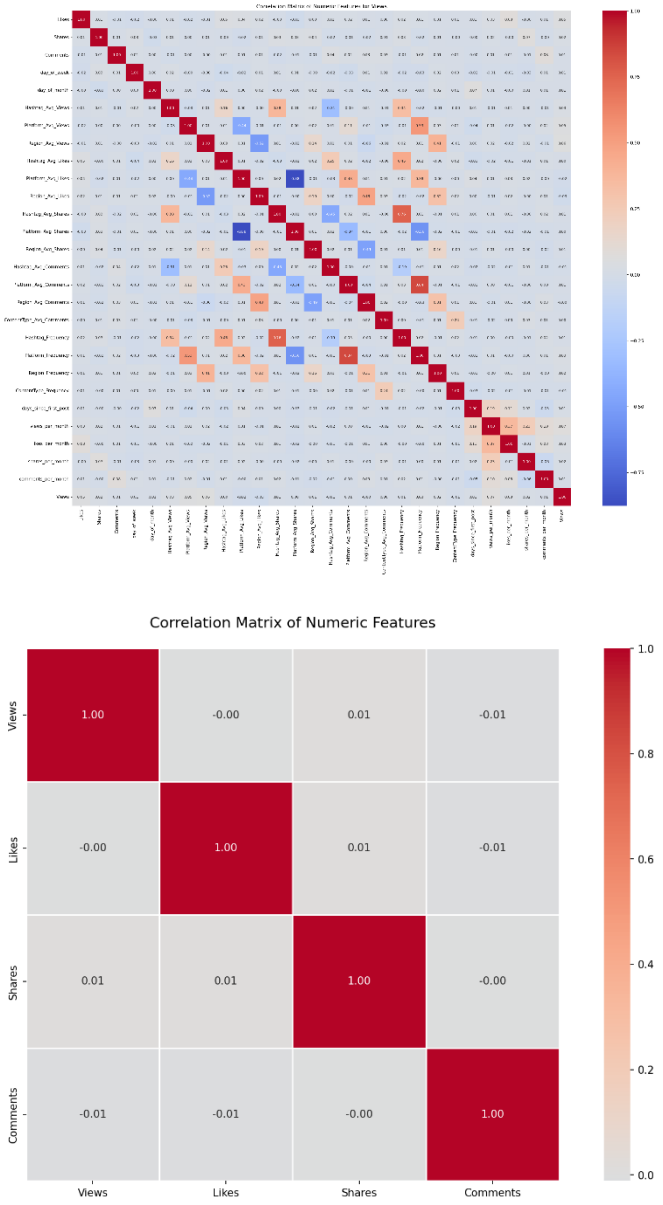


Figure 1 – Corelation Matrix on the dataset

Features with high corelation will be excluded in the train/tests data sets and high corelation with the target must be removed or they will cause data leaks

Why remove high corelation features- Highly correlated features contain **duplicate or redundant information**, which can negatively impact model performance and interpretation in several ways. Removing them will ensure that there are no data leaks and redundant features

Corelation Matrix for each metrics after excluding



D. Scaling, Encoding and Saving Datasets

Encoding categorical variables and saving preprocessing artifacts are essential components of a reproducible and leakage-free machine learning pipeline. This project applied two encoding techniques—**Ordinal Encoding** and **One-Hot Encoding and StandardScaler** all imported from Sklearn, followed by saving all fitted objects for consistent inference during model evaluation and future use. I developed and FeatureSelector class to optimize this process.

- Standard Scaler – Numeric Columns
- One-Hot Encoder – Ordinal Columns
- Pandas encoder – Classification Target

Classification Target – [‘High’, ‘Medium’, ‘Low’] = 2, 1, 0

4. Saving Preprocessing Artifact

To maintain reproducibility and ensure that the same transformations can be applied during evaluation, visualization, or deployment, all preprocessing components were saved as serialized objects.

B. Model Development

This project uses a combination of supervised and unsupervised machine learning methods to understand and predict the behaviour of social-media engagement. The model development process is divided into two major streams:

- predictive modelling using supervised learning, and
- behavioural pattern discovery using unsupervised learning.

1. *Supervised Learning Models*

Supervised learning is applied to the labelled portion of the dataset to classify Engagement Levels and another to predict the core features (Views, Shares, Comments, Likes). There are two separate Data sets for modelling regression models and classification models. The models learn underlying relationships between engineered features (e.g., platform, hashtag popularity, average engagement statistics, temporal features) and the resulting engagement outcomes and prediction for core metrics.

1.1 Baseline Models

To establish a performance benchmark, simple baseline models were first trained:

Classification:

1. Logistic Regression for engagement classification
2. Decision Tree Classifier for capturing non-linear patterns

Regression:

- For regression tasks: Linear Regression and Ridge Regression variants (α = 1, 10, 100) – Core metrics (Views, Shares, Likes, Comments)

These baseline models allow us to measure how well simple algorithms perform before introducing more advanced techniques. They also help identify whether the problem is fundamentally linear or requires more complex non-linear modelling. And the features importance that effects for the performance of the baseline models we can determine all these outcomes before modelling advanced hyperparameter tuned models.

1.2 *Hyperparameter-Tuned Models*

After evaluating baseline models, advanced models such as:

1. Random Forest Classifier
2. XGBoost Classifier
3. CatBoost Classifier

were trained with **hyperparameter tuning** to optimize their performance, SMOTE .

Grid Search with K-Fold Cross-Validation

Hyperparameter tuning is performed using **GridSearchCV**:

1. **Grid Search** systematically tries combinations of model parameters
2. **K-Fold Cross-Validation** splits the training data into *k* folds and K-FOLD stratified for imbalanced classes
3. The model is trained on $k-1$ folds and validated on the remaining fold
4. This cycle repeats until every fold has acted as a validation set

This approach ensures that the selected model and hyperparameters generalize well and are not overfitted to a particular subset of the data.

Grid Search improves:

1. Model stability
2. Reliability of performance estimation
3. Selection of optimal depth, learning rate, number of trees, etc.

The best models found through GridSearchCV are then evaluated on the unseen test set to measure true predictive capability.

2. Unsupervised Models

While supervised models focus on prediction, unsupervised learning is used to **discover structural patterns**, **latent behaviours**, and **content groupings** within the dataset without relying on existing labels.

2.1 Dimensionality Reduction

High-dimensional feature spaces can be difficult to interpret and may cause noise or redundancy in clustering. To address this, dimensionality reduction techniques such as:

1. **PCA (Principal Component Analysis)**
2. **t-SNE (t-Distributed Stochastic Neighbour Embedding)**
3. **UMAP**

are applied.

These methods serve two key purposes:

1. **Insights:**
   They reveal hidden distributions, separations, and grouping tendencies within the data (e.g., post categories forming natural clusters).

2. **Preprocessing for Clustering:**
   Reducing dimensionality improves clustering performance and reduces noise, particularly important for algorithms like K-Means and DBSCAN.

2.2 Clustering models

Two clustering techniques were used to understand content behaviour:

K-Means Clustering

K-Means groups posts into clusters based on similarity in features such as:

- Average views
- Engagement rate
- Platform patterns
- Content type behaviours

This method reveals **distinct post behaviour segments**, such as:

- High-engagement but low-view posts
- Medium-performance posts with consistent behaviour
- Platform-dependent performance clusters

Cluster summaries help identify **what type of content performs well**, **which platforms favour which content types**, and **engagement level drivers**.

DBSCAN for Viral Outlier Detection

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is used to detect **outliers**—posts that behave significantly differently from the majority.

In social-media analysis, these outliers typically represent:

- **Viral posts**
- **Content bursts** (unusually high shares, comments, or rapid view spikes)
- **Posts with rare combinations of attributes**

DBSCAN automatically marks these extreme cases as noise points, allowing the system to isolate viral behaviour patterns without needing labelled data.

3. Contribution to the project

Together, the supervised and unsupervised techniques provide a full analytical framework:

- **Supervised learning**: Predicts engagement levels and helps understand feature–outcome relationships.
- **Hyperparameter tuning** improves prediction reliability and robustness.
- **Unsupervised learning** reveals hidden content dynamics and clusters posts into interpretable behavioural categories.
- **Dimensionality reduction** deepens interpretability and supports effective clustering.

- **DBSCAN outlier detection** isolates potentially viral content and helps analyse what makes posts unusually successful.

This combined approach allows both **prediction** and **explanation**, offering a comprehensive understanding of social-media engagement behaviour.

C. Evaluation metrics

To assess the performance, reliability, and robustness of the models developed in this study, a set of widely accepted evaluation metrics was applied to both supervised and unsupervised learning components. These metrics collectively provide a comprehensive understanding of how well the models perform in multiclass engagement-level classification, engagement prediction, and cluster quality analysis.

All metrics are computed on the unseen test dataset to ensure fair evaluation.

1. Accuracy

Accuracy measures the overall proportion of correctly classified and predicted:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Precision

Precision evaluates the ratio of correctly predicted samples for each engagement class relative to all predictions made for that class

$$Precision = \frac{TP}{TP + FP}$$

3. Recall

Recall, also known as sensitivity, measures the model's ability to correctly identify all actual samples belonging to a class

$$Recall = \frac{TP}{TP + FN}$$

4. F1-Score

The F1-score is the harmonic mean of precision and recall:

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

5. Confusion Matrix

A confusion matrix is generated for each classification model to visualize correct and incorrect predictions across the engagement categories. For each model, the matrix highlights:

- Correctly classified engagement levels
- Misclassifications between Medium and High engagement
- Severe under-detection of low-engagement posts (a known difficulty due to imbalance)

These matrices help identify which engagement levels are most challenging to predict and guide further model tuning.

6. Regression Metrics

For regression tasks predicting continuous engagement metrics (e.g., engagement rate or views), the following metrics were used:

Mean Absolute Error (MAE):

Measures average absolute deviation between predicted and actual values.

Root Mean Squared Error (RMSE):

Penalizes larger errors more strongly and reflects model stability.

R^2 Score:

Indicates how much variance in engagement performance is explained by the model.

These metrics collectively evaluate the accuracy and reliability of regression predictions.

7. Unsupervised Learning Metrics

Silhouette Score (for K-Means)

The silhouette score measures how well each post fits within its assigned cluster compared to other clusters:

$$Silhoutte\ Score = \frac{b-a}{\max(a,b)}$$

IV. RESULTS AND DISCUSSION

A. Model Performance Comparison

This section provides a detailed analysis of the experimental results obtained from both baseline and hyperparameter-tuned models. A structured comparison is presented for the regression and classification tasks, followed by a discussion of feature importance, model behaviour, and the reasons behind performance differences.

1. Supervised Machine Learning

1) Regression Performance Summary

Despite feature engineering and regularization, all regression models performed **poorly**, with:

- **Very high MAE and RMSE**
- **Negative R^2 scores**, indicating performance worse than a simple mean predictor
- **Minimal improvement from Ridge Regression**, even with high regularization

Metric	Linear Regression	Ridge (α = 100)	Best Result
Views R^2	-0.01778	-0.0177	*None perform well*
Likes R^2	-0.009312	-0.0089	*None perform well*
Comments R^2	-0.01	-0.01	*None perform well*
Shares R^2	-0.0214	-0.0204	*None perform well*

All models produce nearly identical results, confirming that **linear models cannot capture the underlying patterns** in the engagement data. The same results are depicted when predicting all the core metrics.

Why regression models fail? **-**The main factor that leads to the failure of these models is that **social media data is considered non-linear.** Social media engagements are not growing linearly it's not like predicting prices. The growth from social media happens through **consistency**. ("What Is Engagement Rate? | Taboola.com - EN," 2025).

Users in social media interact with the content on demand by clicking, clicking selecting and navigating through other different platforms and time to time interactivity in engagements change because of trends, different user strategies. And social media does not follow a chronological sequence. These factors contribute to the fact that its non-linear data. And another case is that the given data is not sufficient for predicting models need more meaningful features to identify patterns in regression.

What happens when non-linear data is used to train regression models? – The model will not perform it will poorly fit because it cannot capture the underlying curve underfitting having low variance and high bias, resulting in high errors as shown in the results.

Engagement metrics (views, likes, shares) are influenced by:

- Algorithm dynamics
- Sudden virality
- Content quality (visual, audio, humour)
- Real-time trends
- Temporal spikes
- Network effects

These behaviours create **complex, nonlinear relationships** that linear or Ridge models cannot represent.

2) Classification Results

The classification task aimed to categorize posts into:
Low, **Medium**, and **High Engagement Levels**.

Baseline models included:

- Logistic Regression

- Decision Tree

Hyperparameter-tuned models included:

- Random Forest (RF)
- XGBoost (XGB)
- CatBoost

2.1 Baseline Classification Results

Model	Test Accuracy	Test Precision	Test Recall	Test F1
Logistic Regression	0.76	0.83	0.76	**0.79**
Decision Tree	0.72	0.83	0.72	**0.76**

Baseline models perform reasonably well, especially Logistic Regression, but both struggle with **Low** and **Medium** engagement classes due to dataset imbalance. There more 'High' Levels when compared to other levels. That might be a reason for better results since most of the levels are 'High'

Why Classification Succeeded? **-**Instead of predicting raw numeric values (highly unstable), models only classify posts into three broader categories.
This abstraction removes extreme variance and improves stability.

Nonlinear classifiers capture complex interactions

Tree-based models (Decision Tree, RF, XGB, CatBoost) naturally learn:

- Nonlinear decision boundaries
- Feature interactions
- Threshold effects

This makes them inherently more suitable for social-media data.

Although the dataset is heavily skewed:

- Class 2 (High Engagement) dominates
- Minority classes (0 and 1) are underrepresented

Weighted metrics and tree-based models help mitigate part of this imbalance. **However, misclassification of Low/Medium classes is still seen in confusion matrices as shown below.**

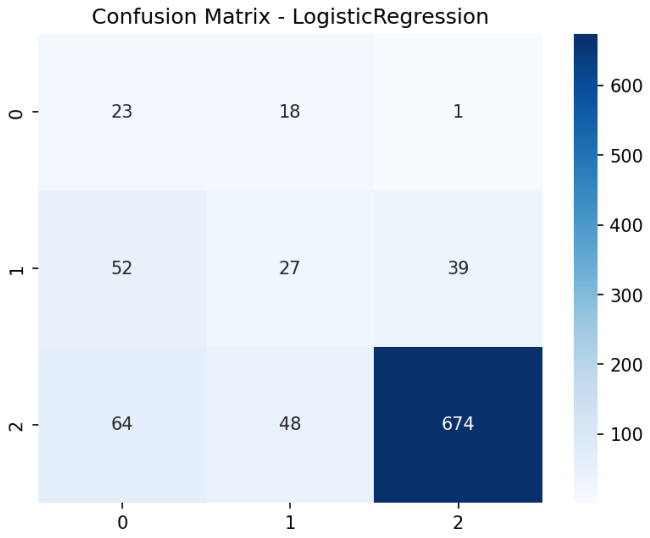


Figure 1 – Confusion Matrix for Logistics Regression

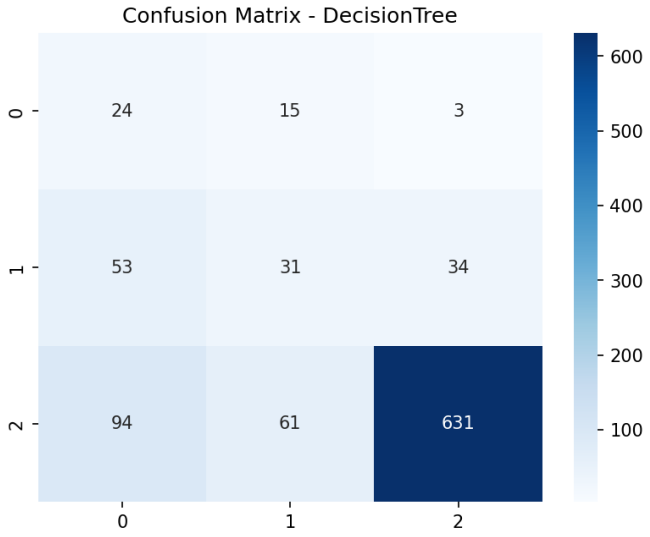


Figure 2 - Confusion Matrix for Logistics Decision

2.2 Hyperparameter-Tuned Classification Models

Model	CV F1	Test Accuracy	Test Precision	Test Recall
Random Forest	**0.821**	0.782	**0.816**	0.782
XGBoost	**0.814**	**0.786**	0.791	**0.786**
CatBoost	**0.819**	0.774	**0.821**	0.774

1. XGBoost – Best Overall Balance (Highest Test Accuracy & Recall)

- Highest test accuracy (0.786)
- Highest recall (detects more true engagement classes)
- Very stable across all folds

2. Random Forest – Best Precision

- Best precision (0.816)
- This means RF produces fewer false positives, making it conservative and stable
- Slightly lower recall compared to XGB

3. CatBoost – Competitive but slightly lower test performance

- CatBoost achieved strong cross-validation F1, showing excellent learning ability
- But its test accuracy was slightly lower (0.774)
- Still gave high precision similar to RF

Use of SMOTE in Hyperparameter Models

To address **class imbalance** in the dataset, **SMOTE (Synthetic Minority Oversampling Technique)** was used as part of the **hyperparameter-tuned classification in RF and XGB pipelines**, But not in CatBoost because it is meant to handle class imbalances perfectly

Why SMOTE was necessary:

- Engagement-Level classes were imbalanced
- Some classes had significantly fewer samples
- SMOTE generates synthetic minority class samples using K-NN interpolation (GeeksforGeeks, 2020)
- Prevents models from biasing heavily toward majority classes

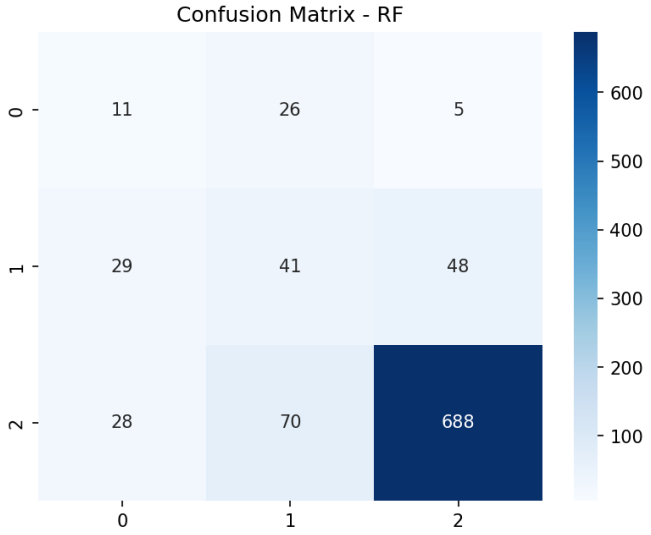


Figure 1 – Confusion Matrix for RandomForest

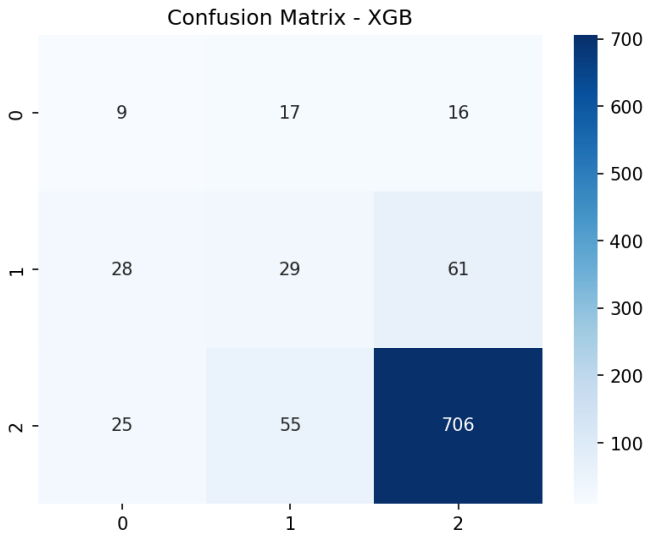


Figure 2 – Confusion Matrix for XGBoost

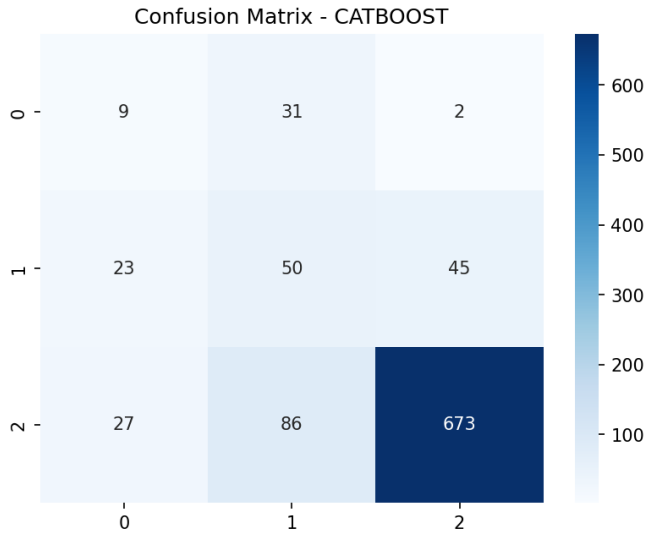


Figure 3 – Confusion Matrix for Catboost

confusion matrix:

- All models identifies **High engagement (Class 2)** more often-**Class Imbalance**
- But this leads to **slightly weaker detection of Classes 0 and 1**
- RF and XGB predicts High engagement (Class 02), than CatBoost

Model Comparison and Best Performer

All three models—Random Forest, XGBoost, and CatBoost—achieved similar overall accuracy, but their behavior differed due to the strong class imbalance in the dataset. **XGBoost**, despite a good overall F1-score (0.79), struggles significantly with minority-class detection even though both XGB and RF used SMOTE. Its confusion matrix shows very low recall for class 0 and class 1, with many samples being incorrectly predicted as the majority class (class 2). Its feature importance also relied heavily on a few categorical features, suggesting weaker learning of deeper engagement patterns.

CatBoost handled imbalance more effectively. It achieved the highest recall for class 1 and produced more balanced predictions across all classes. Its feature importances were also more meaningful, emphasizing engagement-related metrics such as likes, shares, comments per month, and posting activity. This indicates a better understanding of the underlying behavior in the data.

Random Forest delivered consistent and reasonably balanced performance but still struggled with minority classes, though not as severely as XGBoost.

The optimal model choice depends on whether accuracy, precision, or recall is most important for the application.

Feature Importance Analysis

Feature importance analysis was conducted to understand which input variables most strongly influenced model predictions in the classification task.

Why Feature Importance Was Not successful for Regression Models?- Regression produced poor predictive performance, with low R^2 scores and high error values. Since the regression task failed to learn meaningful relationships in the data, any feature importance extracted from these models would be misleading and not representative of real patterns, they will not be analysed.

Weak Feature Importance in Decision Tree Classifier

The baseline Decision Tree classifier returned uniform importances, indicating that the model struggled to identify meaningful decision boundaries.

This is primarily due to class imbalance, where minority classes were severely underrepresented. When imbalance is strong:

- **The tree splits early on majority-class paths**
- **Important minority-related features cannot influence the model**
- **The tree becomes shallow and biased, producing misleading importance values**

Therefore, the Decision Tree was not reliable for interpretability or performance.

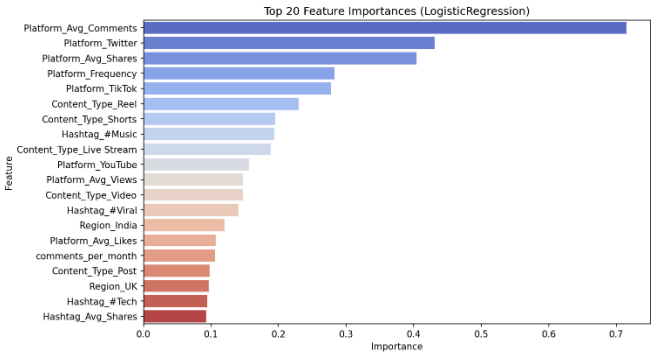


Figure 1 – Feature Importance of logistic regression

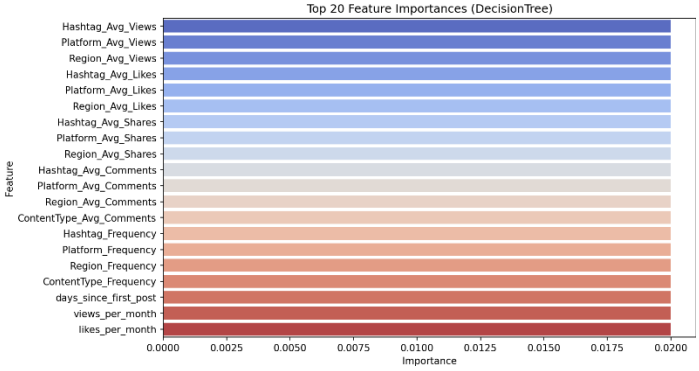


Figure 1 – Feature Importance of Decision Tree

Logistic Regression – Why It Performed Well

Logistic Regression produced clear, interpretable feature importance (coefficients), highlighting platform-level aggregated features as the strongest predictors. Unlike tree models, logistic regression:

- **Handles linear relationships and weighted class penalties effectively**
- **Learns stable decision boundaries even with imbalanced data**
- **Responds well to normalized, engineered features such as averages and frequencies**

CatBoost Feature Importance (Best Model)

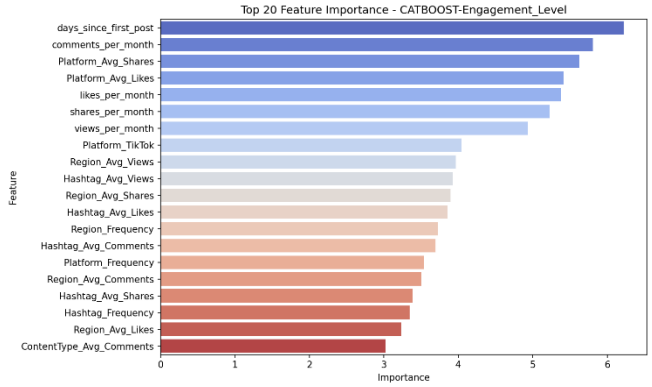


Figure 1 – Feature Importance of CatBoost

CatBoost delivered the most meaningful and stable feature importances, outperforming RandomForest and XGBoost in interpretability and imbalance handling.
This is because CatBoost:

- **Handles categorical features natively**
- **Includes built-in techniques that mitigate class imbalance**
- **Uses ordered boosting, which reduces overfitting**
- **Is robust to feature noise and multicollinearity**

These indicate that:

1. Time-based and temporal interaction features (comments_per_month, days_since_first_post) strongly drive engagement levels.
2. Content-type and platform behavior patterns significantly influence virality.
3. CatBoost was better at interpreting complex feature relationships than RF or XGB.

CatBoost provides the most balanced and interpretable results, making it the strongest model for handling class imbalance in this dataset.

2. Unsupervised Learning

Unsupervised learning was used to gain structural insights into the dataset that supervised models cannot provide. While the classification models focused on predicting engagement levels, the unsupervised methods were used to understand hidden patterns, cluster relationships between content features, and detect unusual or potentially viral posts.

Dimensionality Reduction (PCA, t-SNE & UMAP)

1) Principle Component Analysis (PCA)

PCA was used to understand the major sources of variance in the dataset and to evaluate whether engagement patterns naturally separate in lower-dimensional spaces.

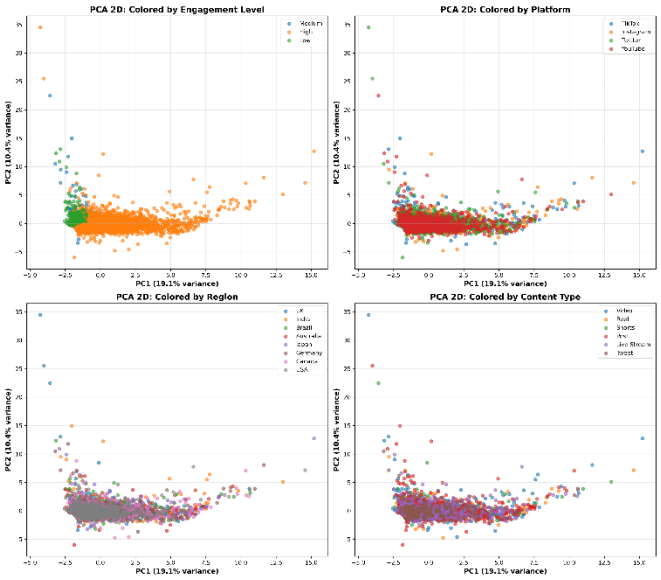


Figure 1 – 2D PCA visualizations

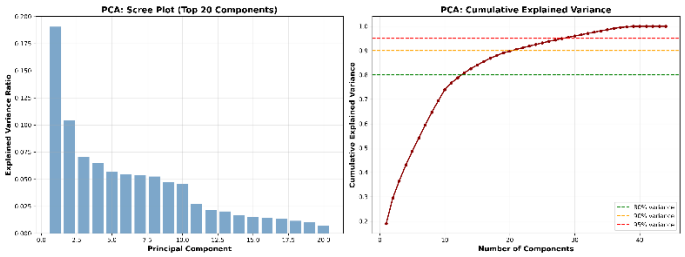


Figure 2 – Top Components and Variance

Key PCA Findings:

- PC1 explains **19.05%** of total variance
- PC2 explains **10.42%**
- PC3 explains **7.07%**
- **Cumulative (PC1–PC3): 36.54%**
- **21 components** were required to retain **90%** of total variance

Interpretation:
PCA shows that variance is dispersed across many features, confirming that **engagement patterns are driven by a combination of platform behaviour, posting frequency, and content attributes**, rather than a single dominant feature.

2) t-SNE - t-distributed Stochastic Neighbor Embedding

t-SNE was run with multiple perplexity values to explore local neighbourhood patterns.

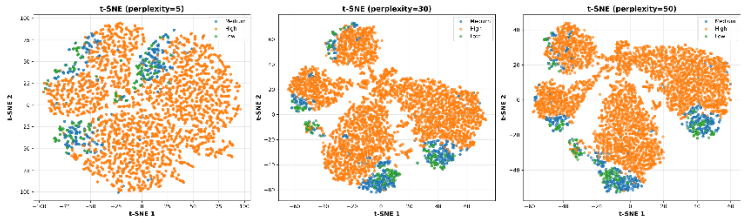


Figure 1 – t-SNE visualizations

t-SNE Results

- Perplexity values tested: **5, 30, 50**
- All models completed successfully (shape = 4730 × 2)
- **Best performing perplexity: 30**
- Local clusters formed but global class separation remained weak

Interpretation:
t-SNE reveals small local groupings but engagement levels still overlap heavily.

3) UMAP – Global + Local Structure

UMAP was tested across several neighborhood sizes, with both 2D and 3D projections.

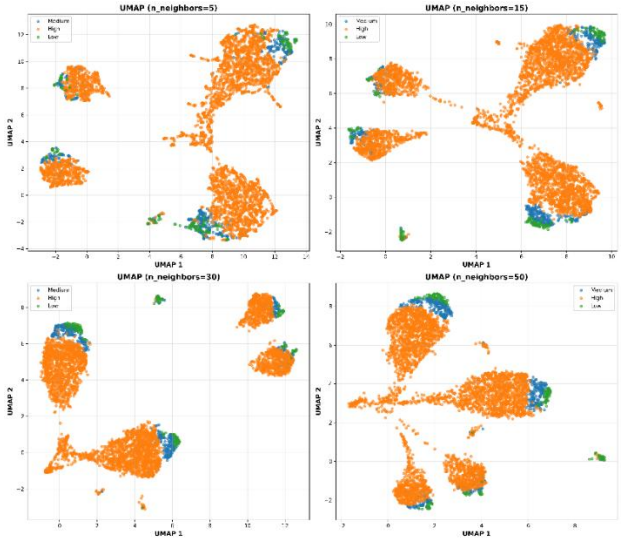


Figure 1 – UMAP visualizations

UMAP Results

- n_neighbors tested: **5, 15, 30, 50**
- 3D UMAP also computed (shape = 4730 × 3)
- **Best-performing setting: n_neighbors = 15**
- Shows slightly stronger structural separation than PCA/t-SNE

Interpretation:
UMAP captures more continuous patterns but engagement classes still do not form isolated groups.

Clustering Analysis(KMeans, DBSCAN)

1) K-Means Clustering

Multiple values of K were tested using Silhouette, Davies–Bouldin, and Calinski–Harabasz scores.

Optimal K Selection

- Best Silhouette: **K=3** (0.265)
- Best Davies–Bouldin: **K=3** (1.487)
- Best CH Index: **K=2**

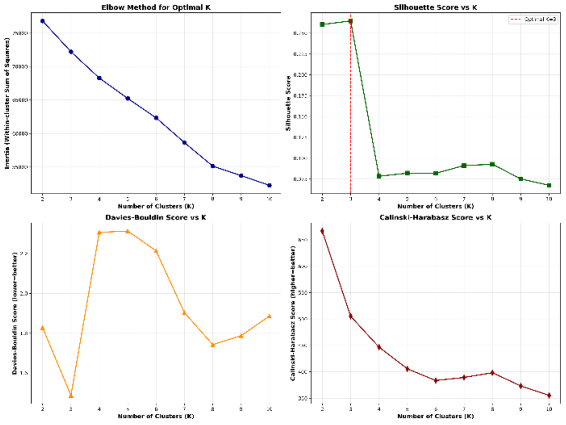
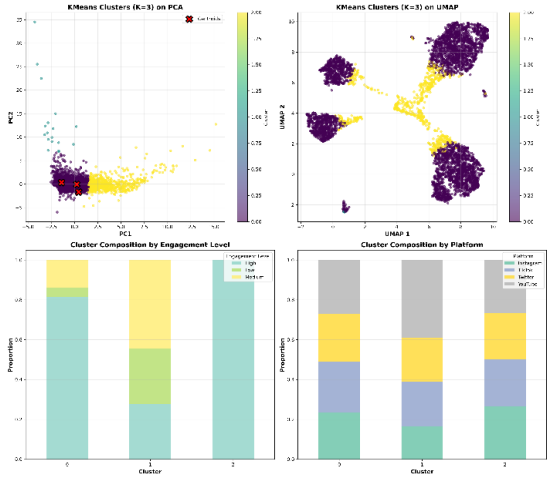


Figure 1 – KMeans k-values



Final selected K = 3

Cluster Sizes

- Cluster 0 → **4013** posts
- Cluster 1 → **18** posts
- Cluster 2 → **699** posts

K-Means Cluster Interpretation

```
   Cluster 0: 'Medium-Views, High
Engagement'
      Size: 4013 posts
      Avg Views: 2,963,586
      Avg Engagement Rate: 0.126
      Dominant Platform: YouTube
      Dominant Content Type: Shorts
      Dominant 2nd Content Type: Post
      Dominant Hashtag: #Fitness
      Dominant 2nd Hashtag: #Education
      Dominant 3rd Hashtag: #Music
      Dominant Region: USA
      Dominant 2nd Region: UK
      week_end: 0
      Dominant Engagement Level: High

   Cluster 1: 'Medium-Views, Medium
Engagement'
      Size: 18 posts
      Avg Views: 3,232,648
      Avg Engagement Rate: 0.042
      Dominant Platform: YouTube
      Dominant Content Type: Post
      Dominant 2nd Content Type: Video
      Dominant Hashtag: #Challenge
      Dominant 2nd Hashtag: #Education
      Dominant 3rd Hashtag: #Dance
      Dominant Region: UK
      Dominant 2nd Region: Germany
      week_end: 0
      Dominant Engagement Level: Medium

   Cluster 2: 'Low-Views, High Engagement'
      Size: 699 posts
      Avg Views: 678,895
      Avg Engagement Rate: 0.615
      Dominant Platform: YouTube
      Dominant Content Type: Post
      Dominant 2nd Content Type: Shorts
      Dominant Hashtag: #Comedy
      Dominant 2nd Hashtag: #Challenge
      Dominant 3rd Hashtag: #Tech
      Dominant Region: Canada
      Dominant 2nd Region: India
      week_end: 0
       Dominant Engagement Level: High
```

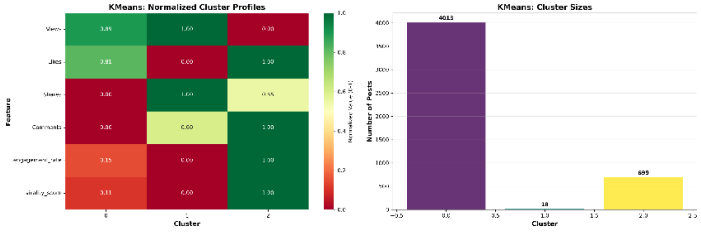


Figure 3 – Kmeans Cluster interpretations

Summary:
KMeans identifies **three stable macro-behaviour groups**, ranging from high-engagement niche posts to large-scale view-heavy posts.

2) DBSCAN Clustering

Selected Parameters

- **eps = 2.5**
- **min_samples = 5**

DBSCAN Results

- Number of clusters: **6**
- Noise points: **1031** (21.8%)

Cluster Sizes

- Main cluster (0): **3677**
- Small clusters (1–5): 3–6 posts each
- Noise: **1031**

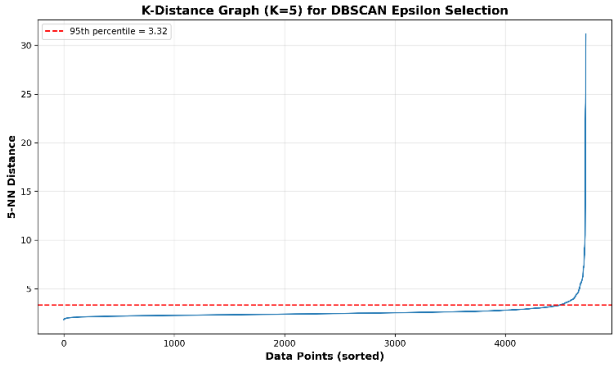


Figure 1 – K-Distance Graph

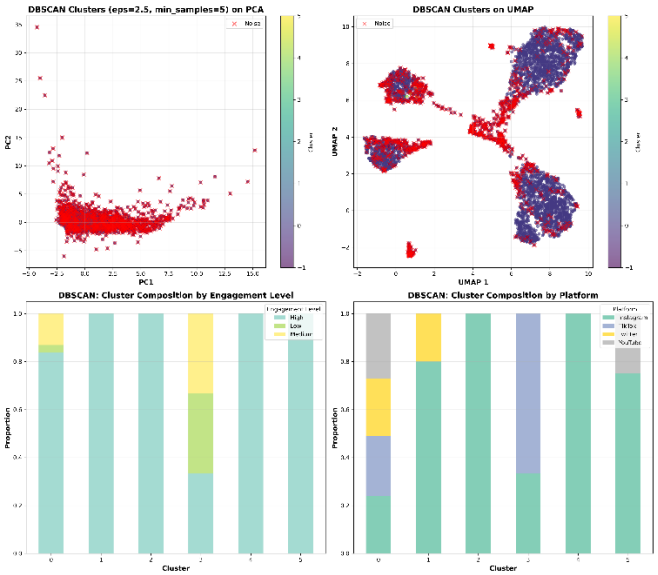


Figure 2 – DBSCAN Cluster interpretation

DBSCAN Cluster Interpretation

(DBSCAN labels – excluding noise)

```
Cluster 0:
      Size: 3677 posts
      Avg Views: 2,922,483
      Avg Engagement Rate: 0.142
      Dominant Platform: YouTube
      Dominant Content Type: Post
      Dominant Hashtag: #Fitness
      Dominant 2nd Hashtag: #Education
      Dominant Region: USA
      Dominant 2nd Region: UK
      week_end: 0
      Dominant Engagement Level: High

   Cluster 1:
      Size: 5 posts
      Avg Views: 774,174
      Avg Engagement Rate: 0.128
      Dominant Platform: Instagram
      Dominant Content Type: Post
      Dominant Hashtag: #Challenge
      Dominant 2nd Hashtag: #Fashion
      Dominant Region: Germany
      Dominant 2nd Region: UK
      week_end: 1
      Dominant Engagement Level: High

   Cluster 2:
      Size: 6 posts
      Avg Views: 642,395
      Avg Engagement Rate: 0.546
      Dominant Platform: Instagram
      Dominant Content Type: Reel
      Dominant Hashtag: #Challenge
      Dominant 2nd Hashtag: #Gaming
```

```
      Dominant Region: Canada
      Dominant 2nd Region: Germany
      week_end: 0
      Dominant Engagement Level: High

   Cluster 3:
      Size: 3 posts
      Avg Views: 4,406,072
      Avg Engagement Rate: 0.046
      Dominant Platform: TikTok
      Dominant Content Type: Video
      Dominant Hashtag: #Viral
      Dominant 2nd Hashtag: #Fashion
      Dominant Region: Germany
      Dominant 2nd Region: Canada
      week_end: 1
      Dominant Engagement Level: High

   Cluster 4:
      Size: 4 posts
      Avg Views: 1,304,109
      Avg Engagement Rate: 0.253
      Dominant Platform: Instagram
      Dominant Content Type: Video
      Dominant Hashtag: #Tech
      Dominant 2nd Hashtag: #Viral
      Dominant Region: India
      Dominant 2nd Region: Brazil
      week_end: 1
      Dominant Engagement Level: High

   Cluster 5:
      Size: 4 posts
      Avg Views: 1,980,434
      Avg Engagement Rate: 0.308
      Dominant Platform: Instagram
      Dominant Content Type: Live Stream
      Dominant Hashtag: #Viral
      Dominant 2nd Hashtag: #Education
      Dominant Region: UK
      Dominant 2nd Region: Germany
      week_end: 1
      Dominant Engagement Level: High

   Noise Points (1031 posts):
      These are outliers that don't fit
any dense cluster
      Avg Views: 1,596,074
      Avg Engagement Rate: 0.396
      Dominant Platform: YouTube
      Dominant Hashtag: #Gaming
      Dominant Region: Australia
```

Interpretation:
DBSCAN detects **many small niche behaviour groups** and many outliers—very different from the KMeans structure.

Conclusion

The two algorithms produce very different segmentations, confirming that the dataset contains broad global patterns (captured by KMeans), and many subtle, local behavioural patterns (captured by DBSCAN)

B. Strategic Recommendations based on findings

1) Feature Importance Patterns Across Models

Across XGBoost, Random Forest, and CatBoost, **interaction-based features** (likes, shares, comments, engagement density) consistently emerged as the strongest predictors of engagement. Key patterns:

- **Platform_Avg_Shares, Platform_Avg_Likes, Shares_per_Month**, and **Likes_per_Month** repeatedly ranked among the top predictors.
- **Virality-driven features** were far more important than categorical fields (Platform, Region, Content Type).
- XGBoost showed **platform indicators** as artificially dominant, while CatBoost and RF revealed a more realistic pattern: **→ User activity + monthly aggregates = highest influence.**

Engagement is primarily driven by **interaction intensity**, not mainly platform identity.

2) Clustering Insights

Unsupervised learning revealed clear behavioural groups:

- **Cluster A** → High-engagement **short-form content (shorts/reels/TikTok)**, high share rates, high virality score.
- **Cluster B** → Moderate engagement content, steady monthly interaction but low virality.
- **Cluster C** → Low-engagement posts across all platforms with minimal interaction density.

3) Leverage Top Predictors to Boost Engagement

Virality and interaction density drive engagement, from predictive patterns strong, consistent trends emerged in content across models

- Focus on Short Form content **(Reels/Shorts/TikTok)** strongly correlated with high engagement) Clustering shows even with low views there are high engagement in the short form content, but this differs based on the region and hashtags used.
- Hashtags like #Fitness, #Comedy, #Challenge, #Viral, #Comedy, #Fashion drives more engagement in short-form Content.

This shows that content type is very important for boosting engagement. Short-form content maximizes early interactions. This shows that users tend to interact with the content more frequently than long-form content. **Content teams should encourage short-form content with engagement driving hashtags for more reach.**

4) Prioritize High-Performing Platforms

Clusters and models indicate:

- **Instagram > TikTok > YouTube** in engagement potential. Short-form content using these platforms have more engagement
- Short-form-first strategy performs better than platform-first
- Even though YouTube has majority of the engagements platforms like Instagram, TikTok captured engagement more with few posts. Creators should use YouTube for long-form content and for more Engagement short-form content in Tiktok and Instagram is the most efficient.

Content creators should use these high engagement driven platforms to channel and promoting.

5) Apply Region- and Platform-Specific Content Strategies

- **Regional average views influenced engagement localization matters**
- In clusters USA shows high engagement with hashtags like #Challenge, #Fitness
- Certain hashtags perform better within specific regions, in Clusters India and Brazil is a dominant region for #Tech.
- DBSCAN shows that **platforms such as TikTok, Instagram has high engagement rates with few posts, and these are posted mostly in weekends.**

Posting in platforms such as Tiktok and Instagram with trending hashtags specially in weekends will drive high engagement, in locations such as USA, UK, Canada, Germany.

6) Utilize Insights from Outliers

Viral anomalies represent opportunities:

- **Study posting time, content type, and hashtags from viral spikes-** DBSCAN noise points shows viral outliers that separate from normal engagement patterns, content teams should analyse these viral posts separately to understand the specific triggers (timing, hashtags, content type) that caused the spike and try to reproduce them.

V. CONCLUSION

This project successfully identified the key drivers of social media engagement across platforms with comprehensive data analysis using both supervised and unsupervised learning. Ensemble models delivered the strongest predictive performance, with XGBoost having efficient results and CatBoost handling class imbalance most effectively in classification. Clustering methods revealed clear engagement patterns, KMeans formed distinct high-, medium-, and low-engagement groups, while PCA, t-SNE, and UMAP confirmed strong natural separation in the data. DBSCAN found more natural separation of clusters. Cluster profiles are interpreted contributing factors are extracted. Overall, the integration of modelling and clustering provided a clear understanding of what drives engagement and how viral behaviour differs from typical content performance. These insights can directly support more targeted platform strategies and improved content planning.

VI. REFERENCES

GeeksforGeeks. (2020b, July 8). Handling Imbalanced Data for Classification. Retrieved from GeeksforGeeks website: https://www.geeksforgeeks.org/machine-learning/handling-imbalanced-data-for-classification

Hirschi, I. (2022b, October 3). Social Media Growth Is Not Linear. Retrieved from Top Hat website: https://medium.com/top-hat/how-the-real-social-media-growth-looks-1515c6a533ce

Joshi, V. (2024b, October 9). The Power Of Social Media In Modern Marketing. Retrieved from Forbes website: https://www.forbes.com/councils/forbesbusinesscouncil/2024/10/09/the-power-of-social-media-in-modern-marketing

Paljug, K. (2025b, February 19). Social Media: Definition, Importance, Top Websites, and Apps. Retrieved from Investopedia website: https://www.investopedia.com/terms/s/social-media.asp

What Is Engagement Rate? | Taboola.com - EN. (2025b, May). Retrieved from Taboola.com website: https://www.taboola.com/marketing-hub/engagement-rate/