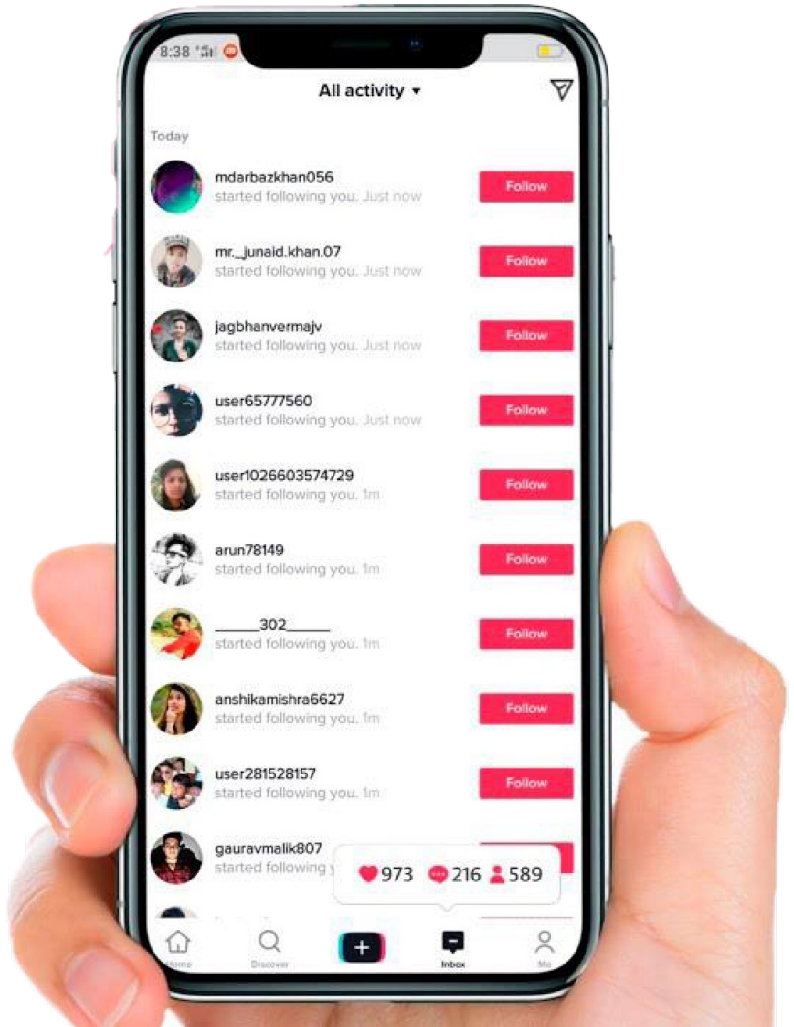
# Data Mining on TikTok
# Video and User Engagement

Jeffrey Huang (jeffretirever27@gmail.com)

# Table of Contents


- Introduction
- Data
- Method
- Analysis and Results
  - Q1 Classify Claim Shorts
  - Q2 Clustering Metrics
  - Q3 Topic Shareability
- Conclusion

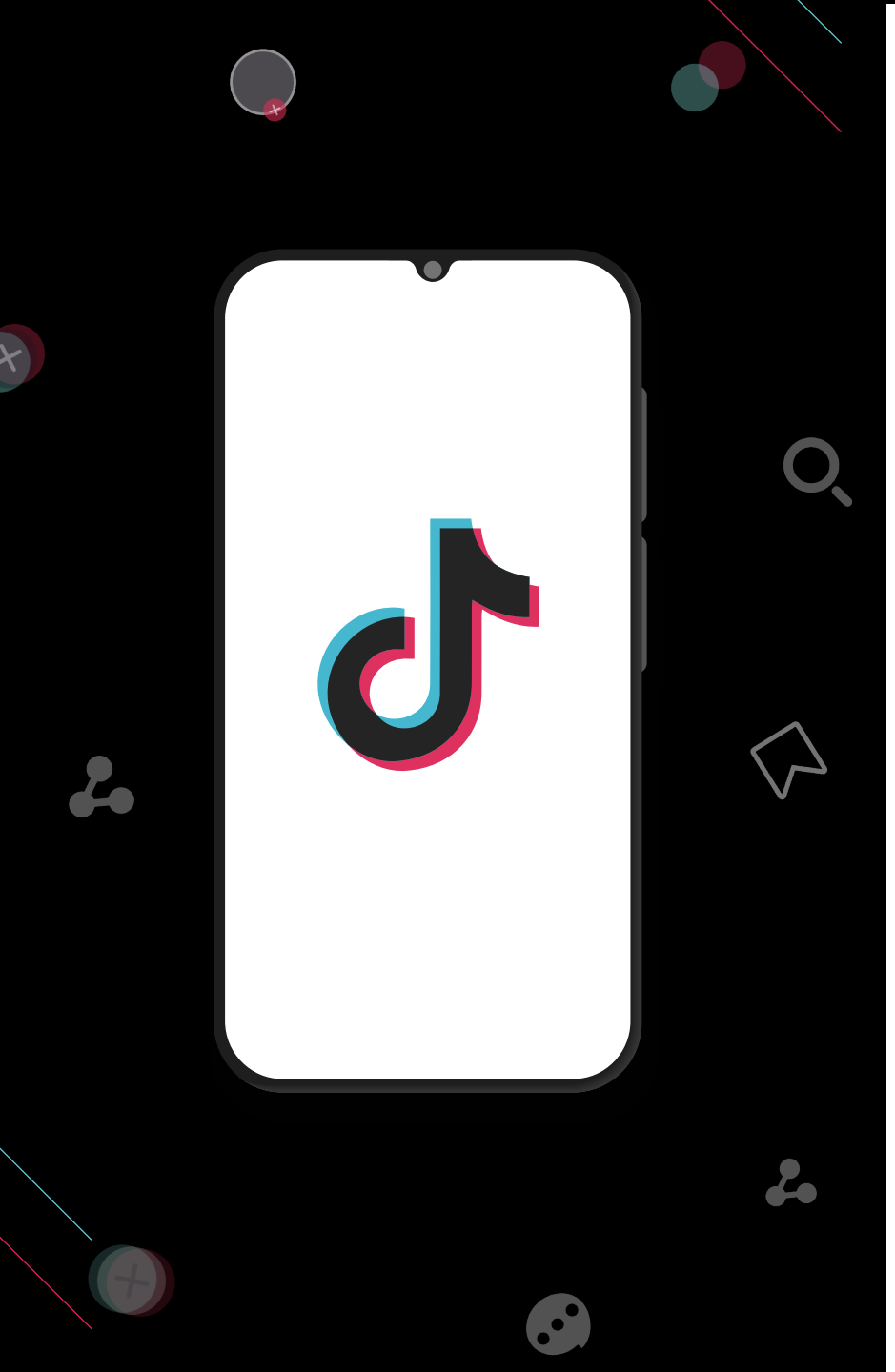
# Introduction

# Introduction

This project proposes a comprehensive data mining study on TikTok to optimize user engagement and ensure the credibility of content. Utilizing a dataset of 19,383 TikTok videos, the study will employ Ensemble Learning, Logistic Regression, EDA, K-means and NLP techniques to predict misinformation, cluster videos by engagement, and analyze content for shareability. This research is motivated by the need for enhanced content personalization, targeted advertising, and responsible platform governance, with implications for broader social media interaction studies.

# The Major Problem
# Misinformation

Misinformation on social media is a critical issue in the digital age, as it shapes public discourse and trust. This project addresses the urgent need to discern fact from fiction by developing algorithms that automatically verify the truthfulness of claims in TikTok videos. Utilizing machine learning models like random forests and boosting, the project seeks to enhance content credibility and uphold platform integrity, ensuring users receive reliable information.



# 3 Research Questions and Strategy

**How can we classify claim shorts or opinion shorts based on data?**

Predicting misinformation is key to a credible platform, using several classification model and Ensemble Learning like Random Forest and Gradient Boosting for robust classification.

**How many types/groups of video based on user engagement?**

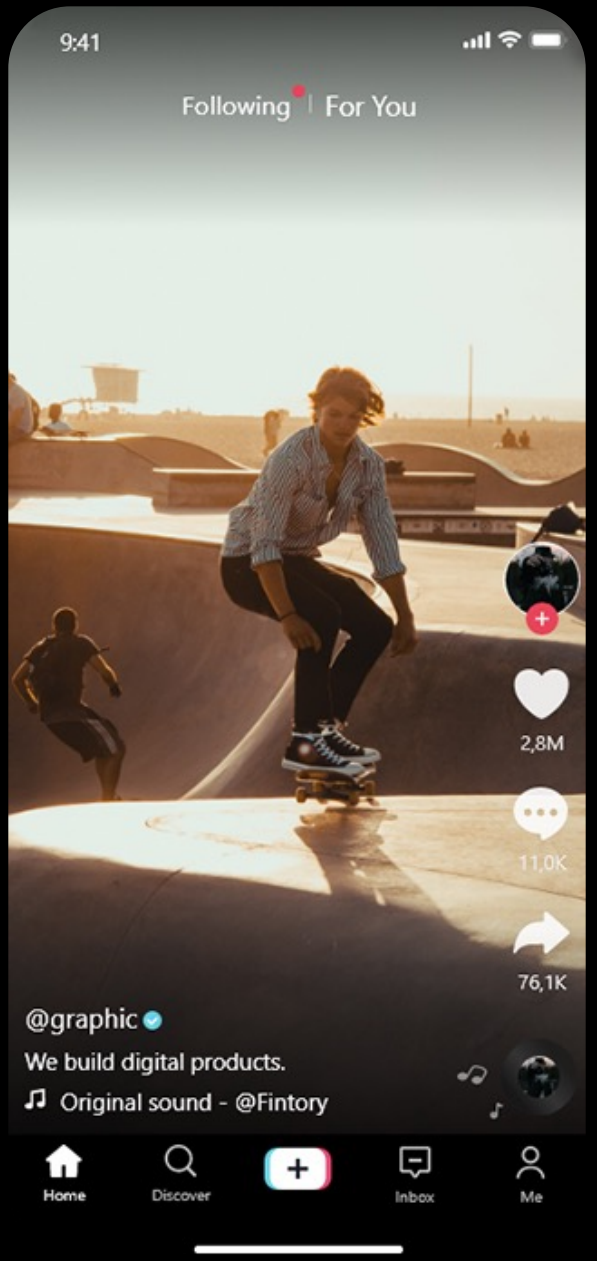
Grouping videos by engagement informs advertising and content strategy, using unsupervised learning like K-means to group similar interaction patterns.

**What topic of shorts make the most shareability?**

Natural Language Processing methods such as topic modeling algorithms and sentiment analysis can be employed for this dual purpose.

# Top 8 Challenges

1. Analyzing and verifying a large dataset with diverse data types and metrics.
2. Ensuring data integrity and preprocessing accuracy for reliable model training.
3. Addressing high multicollinearity among user engagement metrics.
4. Creating accurate classification models for a multifaceted classification problem.
5. Training models to classify content without overfitting or bias.
6. Distinguishing factual claims in TikTok videos to combat misinformation.
7. Topic modeling with inadequate text per data point and the excessive number of topics and data points to model.
8. Adapting findings to broader social media platforms and varied content types.



# Data

# Data Source

The dataset is sourced from **Kaggle**, accessible through the following link: https://www.kaggle.com/datasets/yakhyojon/tiktok/data

The dataset originally contains **19,382** observations and **12** variables. Each row represents a published TikTok video that includes a claim or an opinion. The columns have various data types like integers, floats, and objects and include metrics like view count, like count, and the claim status of a video.

| Column Name | Type | Description |
|---|---|---|
| # | int | TikTok assigned number for video with claim/opinion. |
| claim status | obj | Whether the published video has been identified as an “opinion” or a “claim.” |
| video id | int | Random identifying number assigned to video. |
| video duration sec | int | Video duration in seconds. |
| video transcription text | obj | Transcribed text of the words spoken in the published video. |
| verified status | obj | The status of the user who published the video in terms of their verification. |
| author ban status | obj | The status of the user who published the video in terms of their permissions. |
| video view count | float | Total number of times the published video has been viewed. |
| video like count | float | Total number of times the published video has been liked by other users. |
| video share count | float | Total number of times the published video has been shared by other users. |
| video download count | float | Total number of times the published video has been downloaded by other users. |
| video comment count | float | Total number of comments on the published video. |

**Note:** regarding claim status, an “opinion” refers to an individual’s or group’s personal belief or thought. A “claim” refers to information that is either unsourced or from an unverified source.

# Data Inspection

The first five rows of the dataset are displayed for an initial glance:

| # | claim status | video id | video duration sec | video transcription text | verified status | author ban status | video view count | video like count | video share count | video download count | video comment count |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | claim | 7017666017 | 59 | someone shared with me that drone ... | not verified | under review | 343296 | 19425 | 241 | 1 | 0 |
| 1 | claim | 4014381136 | 32 | someone shared with me that there are ... | not verified | active | 140877 | 77355 | 19034 | 1161 | 684 |
| 2 | claim | 9859838091 | 31 | someone shared with me that American industrialist…… | not verified | active | 902185 | 97690 | 2858 | 833 | 329 |
| 3 | claim | 1866847991 | 25 | someone shared with me that the metro ... | not verified | active | 437506 | 239954 | 34812 | 1234 | 584 |
| 4 | claim | 7105231098 | 19 | someone shared with me that the number ... | not verified | active | 56167 | 34987 | 4110 | 547 | 152 |

**Note:** regarding video transcription text, the value presented in the preceding table has been streamlined.

# Data Cleaning:
# Missing and Duplicates

**Missing Values:**

The last **298** rows are all missing values.

**Handel Method:**

Removed the last 298 rows.

**Duplicate Values:**

**0** duplicate values.

**Handel Method:**

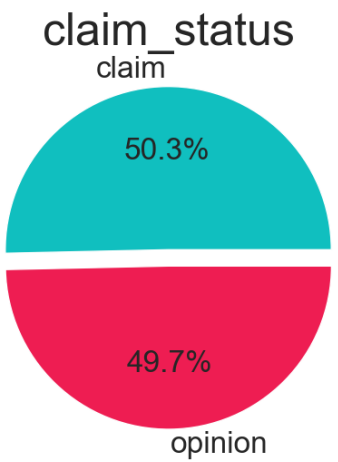
N/A

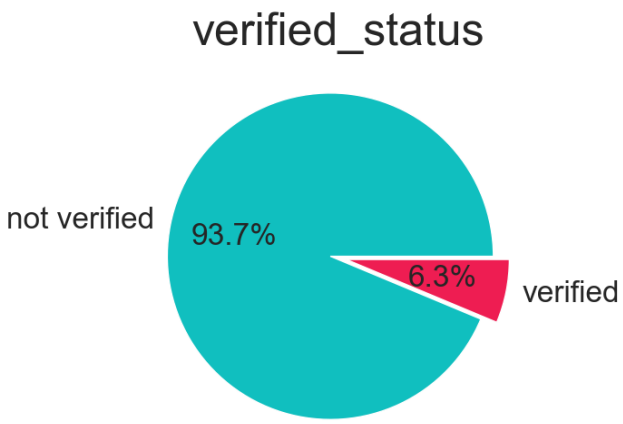
| # | claim status | video id | video duration sec | video transcription text | verified status | author ban status | video view count | video like count | video share count | video download count | video comment count |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 19080 | opinion | 1492320297 | 49 | in our opinion the earth holds about 11 quintillion pounds of air | not verified | active | 6067 | 423 | 81 | 8 | 2 |
| 19081 | opinion | 9841347807 | 23 | in our opinion the queens in ant colonies live for around 30 years | not verified | active | 2973 | 820 | 70 | 3 | 0 |
| 19082 | opinion | 8024379946 | 50 | in our opinion the moon is moving away from the earth | not verified | active | 734 | 102 | 7 | 2 | 1 |
| 19083 | opinion | 7425795014 | 8 | in our opinion lightning strikes somewhere on earth about 100 times every second | not verified | active | 3394 | 655 | 123 | 11 | 4 |
| 19084 | opinion | 4094655375 | 58 | in our opinion a pineapple plant can only produce one pineapple a year | not verified | active | 5034 | 815 | 281 | 11 | 1 |

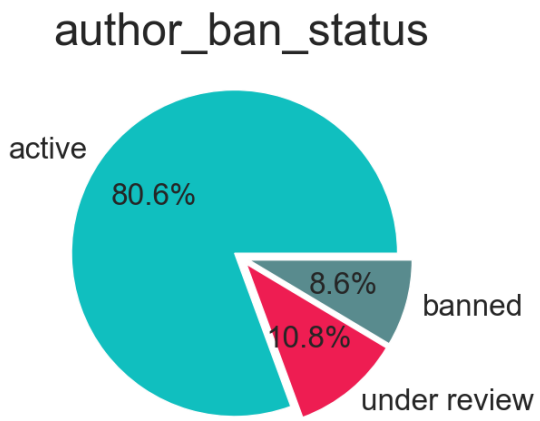
The updated dataset contains **19,084** observations and **12** variables with **0** missing or duplicates values.

# General Exploratory Data Analysis (EDA):
# Categorical Data

| Column Name | Type | Unique Value |
|---|---|---|
| claim status | obj | “opinion” and “claim.” |
| verified status | obj | 'not verified’ and 'verified' |
| author ban status | obj | 'under review’, 'active’ and 'banned' |



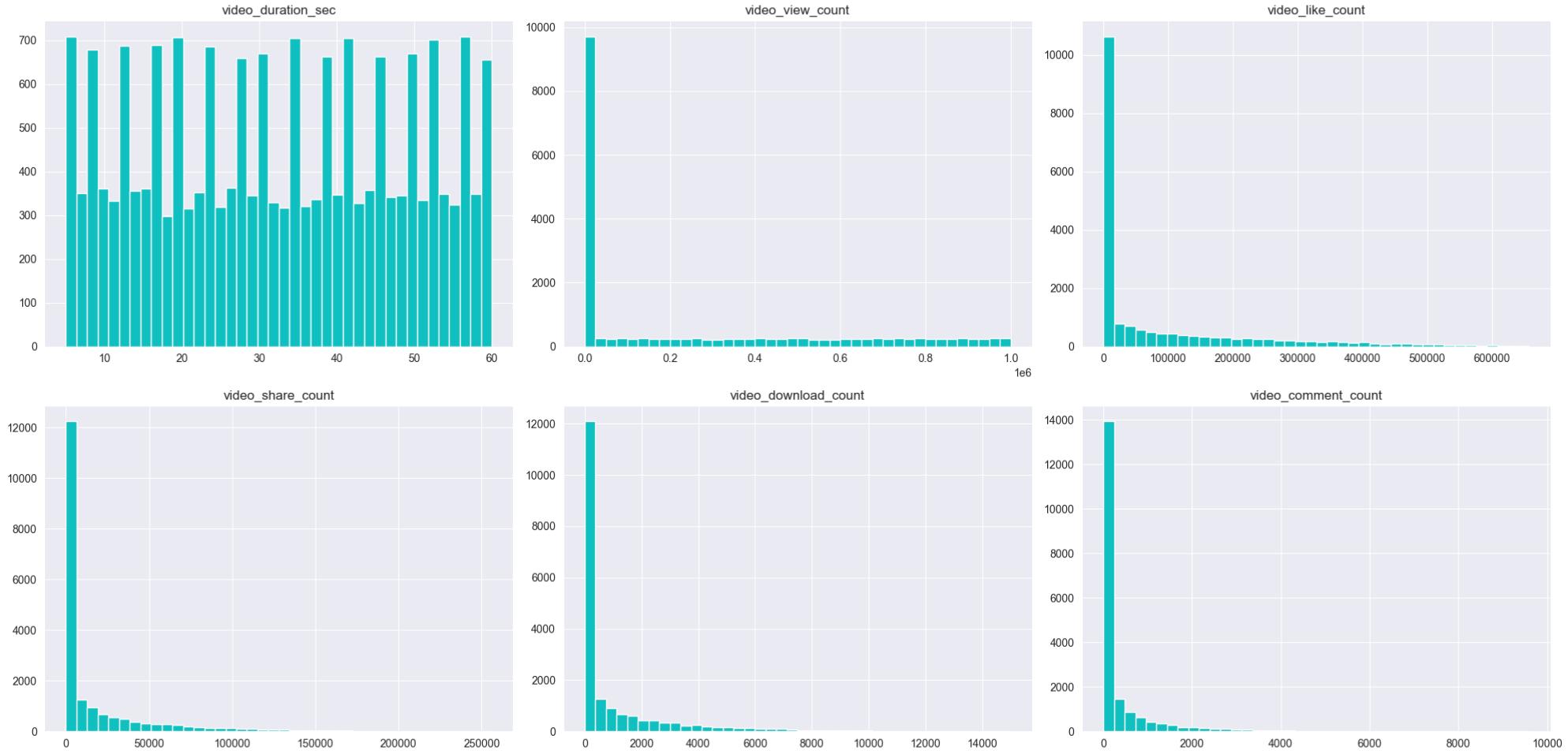




The number of rows where claim status is 'opinion': 9476 which is 49.65%.
The number of rows where claim status is 'claim': 9608 which is 50.35%.
The number of rows where verified status is 'not verified': 17884 which is 93.71%.
The number of rows where verified status is 'verified': 1200 which is 6.29%.

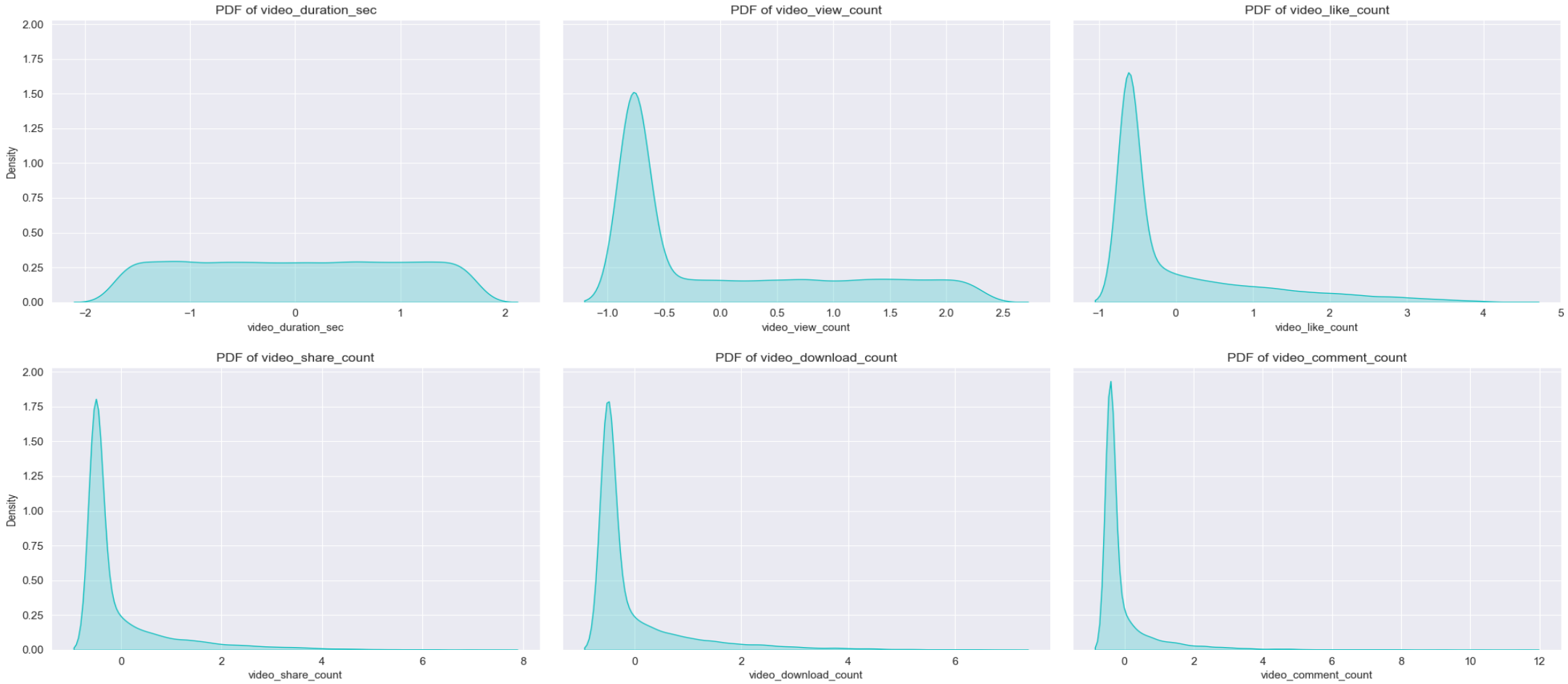
The number of rows where author is 'active': 15383 which is 80.61%.
The number of rows where author is 'banned': 1635 which is 8.57%.
The number of rows where author is 'under review': 2066 which is 10.83%.

# General Exploratory Data Analysis (EDA):
# Numerical Data



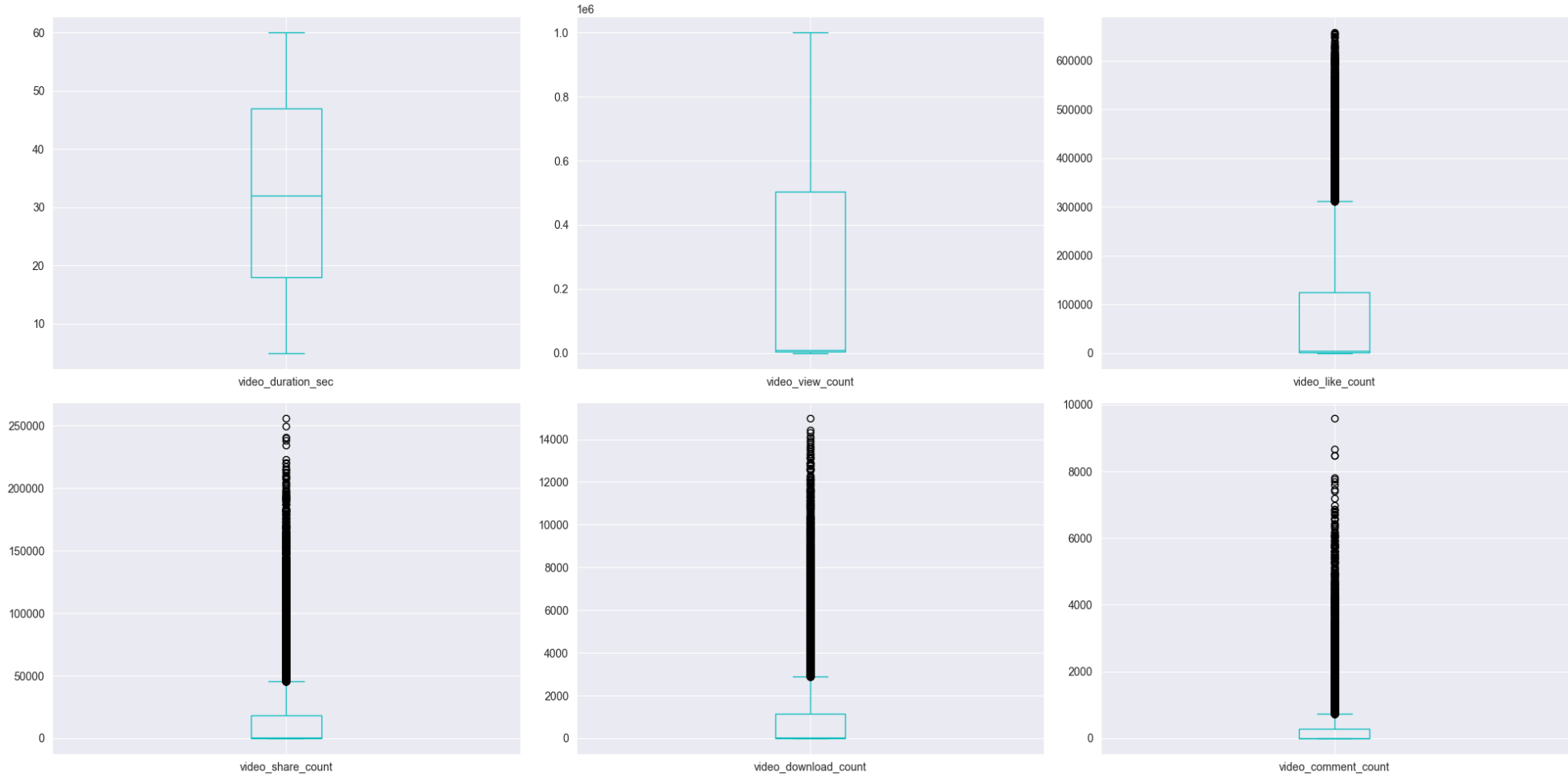

- **Video Duration**: It appears that the videos have a uniform distribution of lengths..
- **Engagement Metrics**: All these metrics display a heavy right-skewed distribution, suggesting most videos have low engagement.

# General Exploratory Data Analysis (EDA): Numerical Data




- **Video Duration**: It appears that the videos have a uniform distribution of lengths..
- **Engagement Metrics**: All these metrics display a heavy right-skewed distribution, suggesting most videos have low engagement.

# General Exploratory Data Analysis (EDA):
# Numerical Data




- **Video Duration**: There are no extreme outlier, and it is uniformly distributed.
- **Engagement Metrics**: There are several extreme outliers that indicate some videos have exceptionally high counts compared to the rest.

# Method

# 3 Research Questions and Strategy

**How can we classify claim or opinion shorts?**

Predicting misinformation is key to a credible platform, using several classification model and Ensemble Learning like Random Forest and Gradient Boosting for robust classification.

**How many groups of video based on user engagement?**

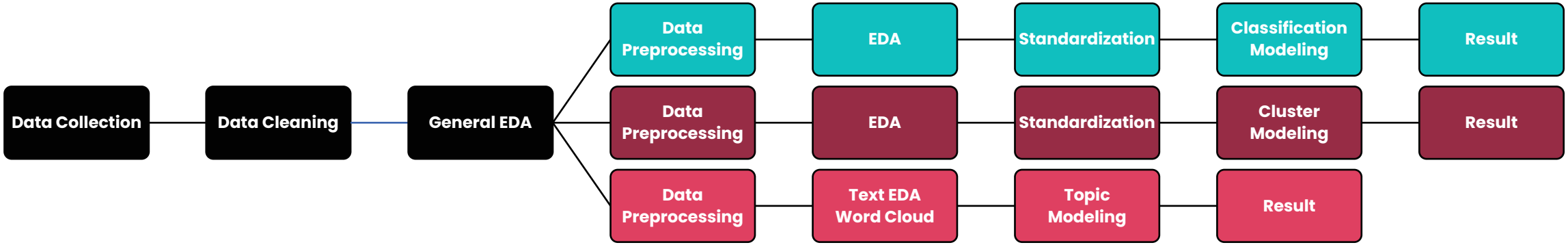
Grouping videos by engagement informs advertising and content strategy, using unsupervised learning like K-means to group similar interaction patterns.

**What topic of shorts make the most shareability?**

Natural Language Processing methods such as topic modeling algorithms and sentiment analysis can be employed for this dual purpose.

# Data Mining Process

- How can we classify claim shorts or opinion shorts based on data?
- How many types/groups of video based on user engagement?
- What topic of shorts make the most shareability?



# Q1. Classify Claim

- **Question**

  How can we classify claim shorts or opinion shorts based on data?

- **Solution**

  Use several classification model Random Forest and Gradient Boosting for robust classification.

- **Basic Feature Selection**

  claim status, video duration, verified status, author status, video views, likes, shares, downloads, comments.

- **Visualization**

  pie charts, cross table, pair plots, KDE plots, correlation heat map, histogram, and line chart.

- **Models**

  Random Forest, Gradient Boosting, KNN, LDA, Logistic Regression, Decision Tree.

- **Challenges**

  (1) Multicollinearity between features. (2) Hyperparameter tuning with cross validation. (3) Avoid overfitting.

- **Process**

# Q2. Clustering Metrics

- **Question**

  How many groups of video based on user engagement?
- **Solution**

  Use unsupervised learning like K-means to group similar interaction patterns.
- **Basic Feature Selection**

  video views, likes, shares, downloads, comments.
- **Visualization**

  Elbow diagram.
- **Models**

  K-means.
- **Challenges**

  (1) Find the best number of clusters – K. (2) Must run for amount of iteration, or it would produce a suboptimal result.
- **Process**

# Q3. Topic Shareability

- **Question**

  What topic of shorts make the most shareability?
- **Solution**

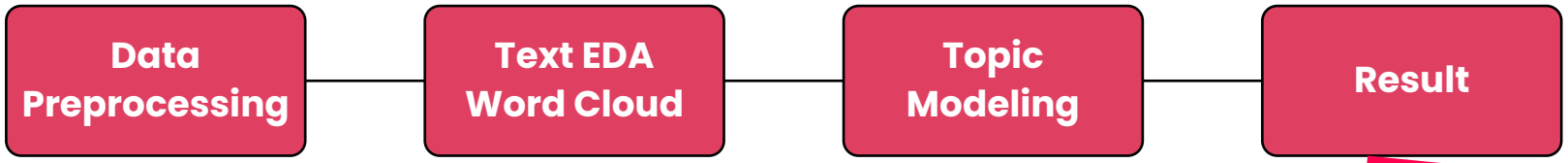
  Use NLP methods such as Latent Dirichlet Allocation and sentiment analysis to cluster topics.
- **Basic Feature Selection**

  video transcription text, video shares, claim status, verified status, author status.
- **Visualization**

  Word cloud , sentiment analysis, and histogram.
- **Models**

  Latent Dirichlet Allocation and Sentiment Intensity Analyzer.
- **Challenges**

  Topic modeling with inadequate text per data point and the excessive number of topics and data points to model.
- **Process**



# Analysis and Results
# Q1 Classify Claim

# Q1. Classify Claim

## Data Preprocessing

- **Feature Selection**

  claim status, video duration, verified status, author status, video views, likes, shares, downloads, comments.

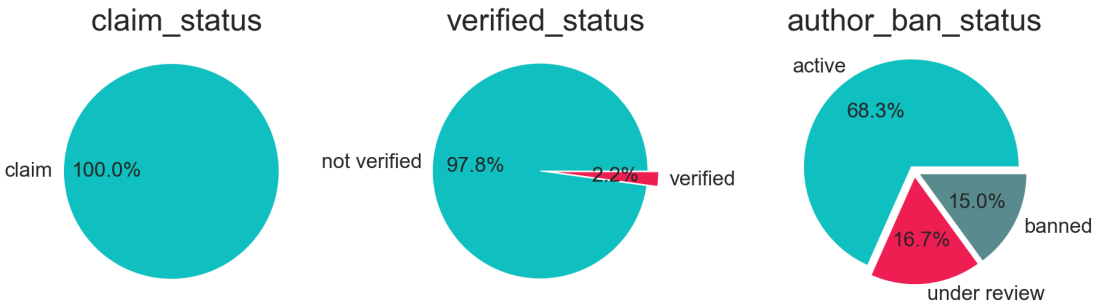
- **Categorical Data Encoding**
  - Convert binary categorical column's value into 0 and 1.
  - Convert trinary categorical column's value into 0 and 1 and using **One-hot encoder** to create 3 columns.

| claim status | video duration sec | verified status | video view count | video like count | video share count | video download count | video comment count | author active | author banned | author under review |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 59 | 0 | 343296 | 19425 | 241 | 1 | 0 | 0 | 0 | 1 |
| 1 | 32 | 0 | 140877 | 77355 | 19034 | 1161 | 684 | 1 | 0 | 0 |
| 1 | 31 | 0 | 902185 | 97690 | 2858 | 833 | 329 | 1 | 0 | 0 |
| 1 | 25 | 0 | 437506 | 239954 | 34812 | 1234 | 584 | 1 | 0 | 0 |
| 1 | 19 | 0 | 56167 | 34987 | 4110 | 547 | 152 | 1 | 0 | 0 |

# Q1. Classify Claim

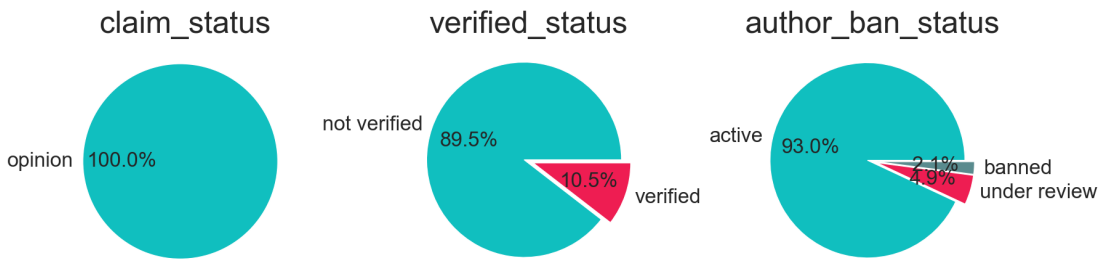
## EDA: Categorical Data

### Pie Chart - Claim Group




- verified status pie chart indicates that within the claims, most of the shorts are not verified.
- author ban status is more diverse than verified status.

### Pie Chart - Opinion Group




- verified status pie chart indicates that within the opinion, not verified status still take a large portion.
- author ban status shows that within the opinion, most of the shorts are from active users.

# Q1. Classify Claim

## EDA: Categorical Data

### Cross Table - Claim and Verification

| \verification status<br>claim status | Not Verified | Verified |
|---|---|---|
| **Opinion** | 8485 | 991 |
| **Claim** | 9399 | 209 |

- Most of both claims and opinions are not verified, but the proportion of opinions that are verified is noticeably higher.

### Cross Table – Claim and Author Status

| \verification status<br>claim status | active | banned | under review |
|---|---|---|---|
| **Opinion** | 8817 | 196 | 463 |
| **Claim** | 6566 | 1439 | 1603 |

- **Opinions** seem to have a much lower rate of authors being banned or under review, indicating potentially less controversy or a different standard of review.

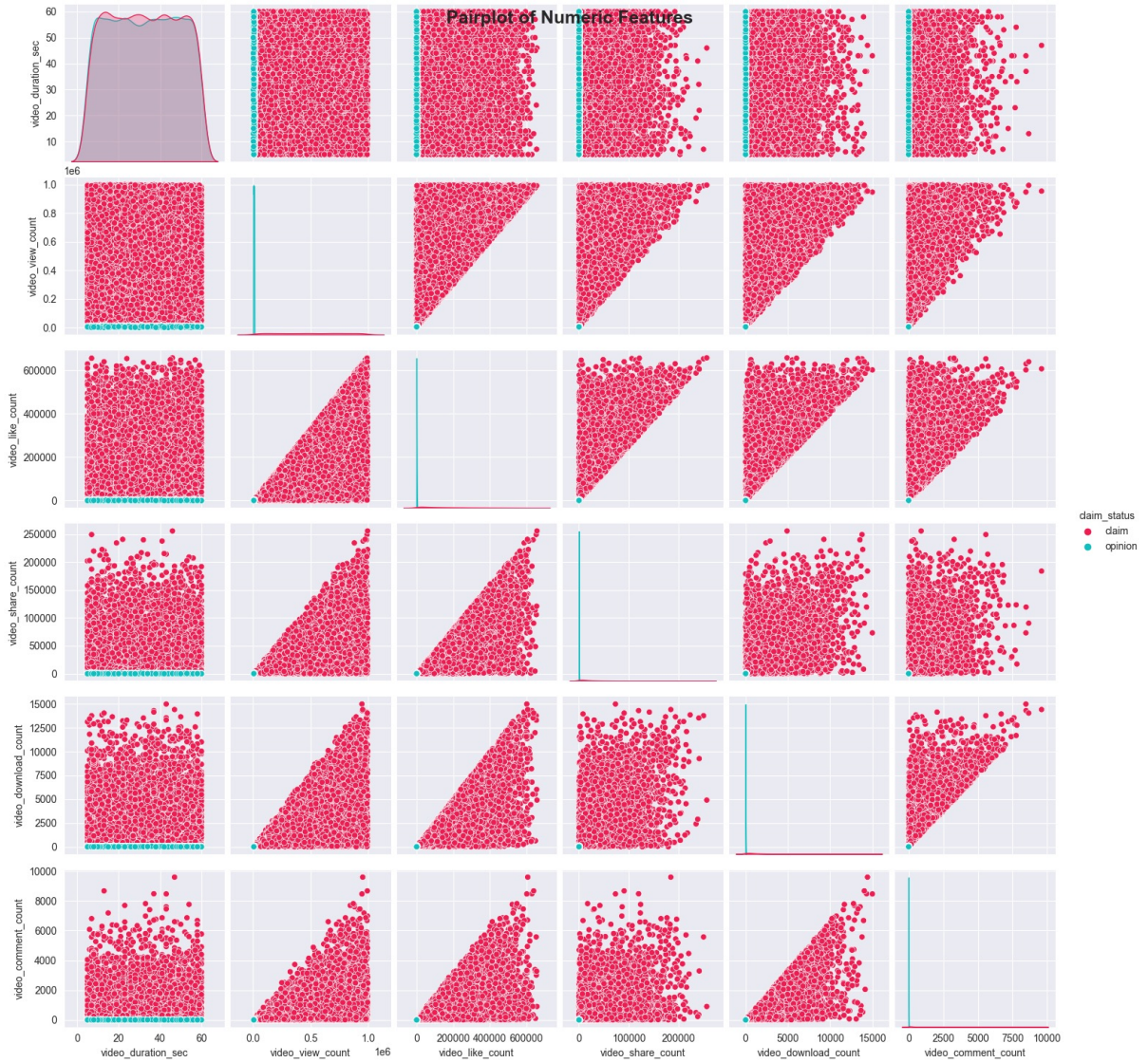
# Q1. Classify Claim

## EDA: Numerical Data

The **video duration** does not show a clear distinction between 'claim' and 'opinion' statuses, indicating that the duration of the video may not be a strong differentiator between these two categories.

***Opinion*** has a lower performance in engagement constantly compared to ***Claim***, and ***Claim*** has a higher variance than ***Opinion***.

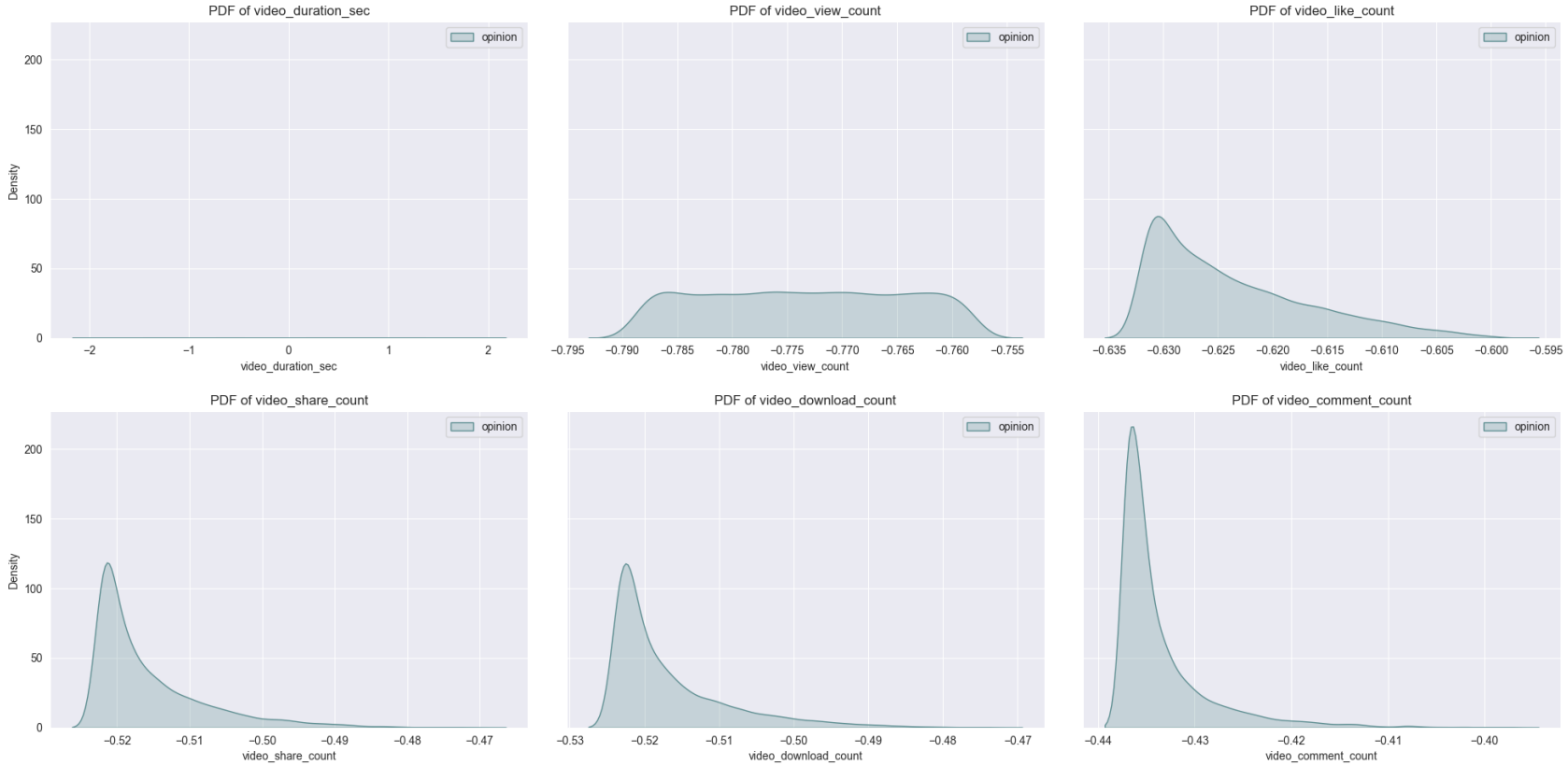
Some variables show potential positive correlations with each other, such as views, likes, shares, and comments. This is expected as videos with more views tend to have more likes, shares, and comments.



# Q1. Classify Claim

## EDA: Numerical Data

### Opinion Kernel Density Estimation

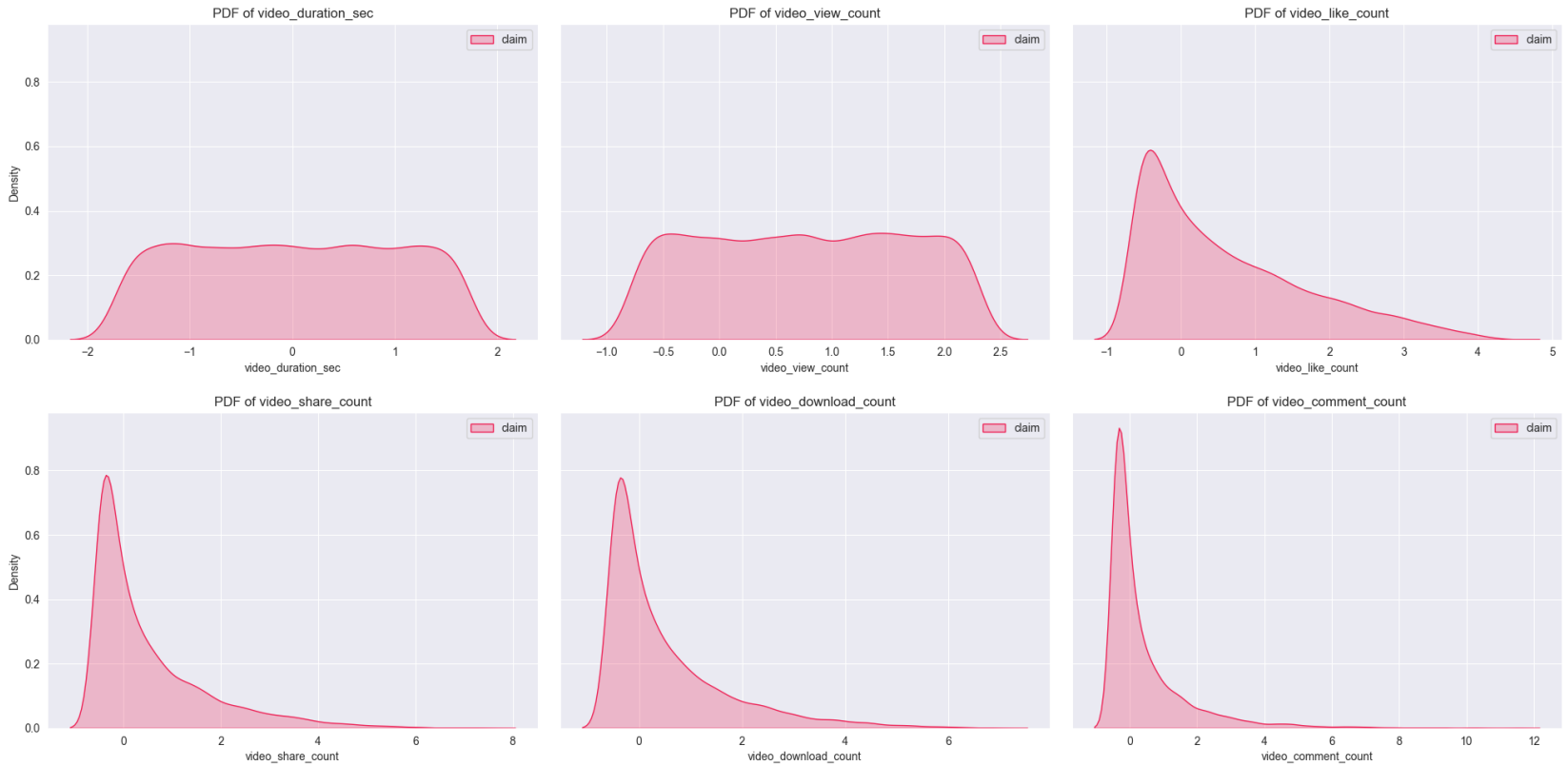



- Most engagement variables appear to have a right-skewed distribution. There are pronounced peaks near zero for most features, which suggests a high concentration of data with low values.
- The distribution of video duration seems less skewed compared to the other metrics.

# Q1. Classify Claim

## EDA: Numerical Data

### Claim Kernel Density Estimation




- While both claim and opinion categories show right-skewed distributions for engagement metrics the claim exhibits slightly broader distributions with more pronounced tails.
- This suggests that while most claim videos have low engagement like opinion videos, there is a subset of claim content that achieves significantly higher engagement.

# Q1. Classify Claim

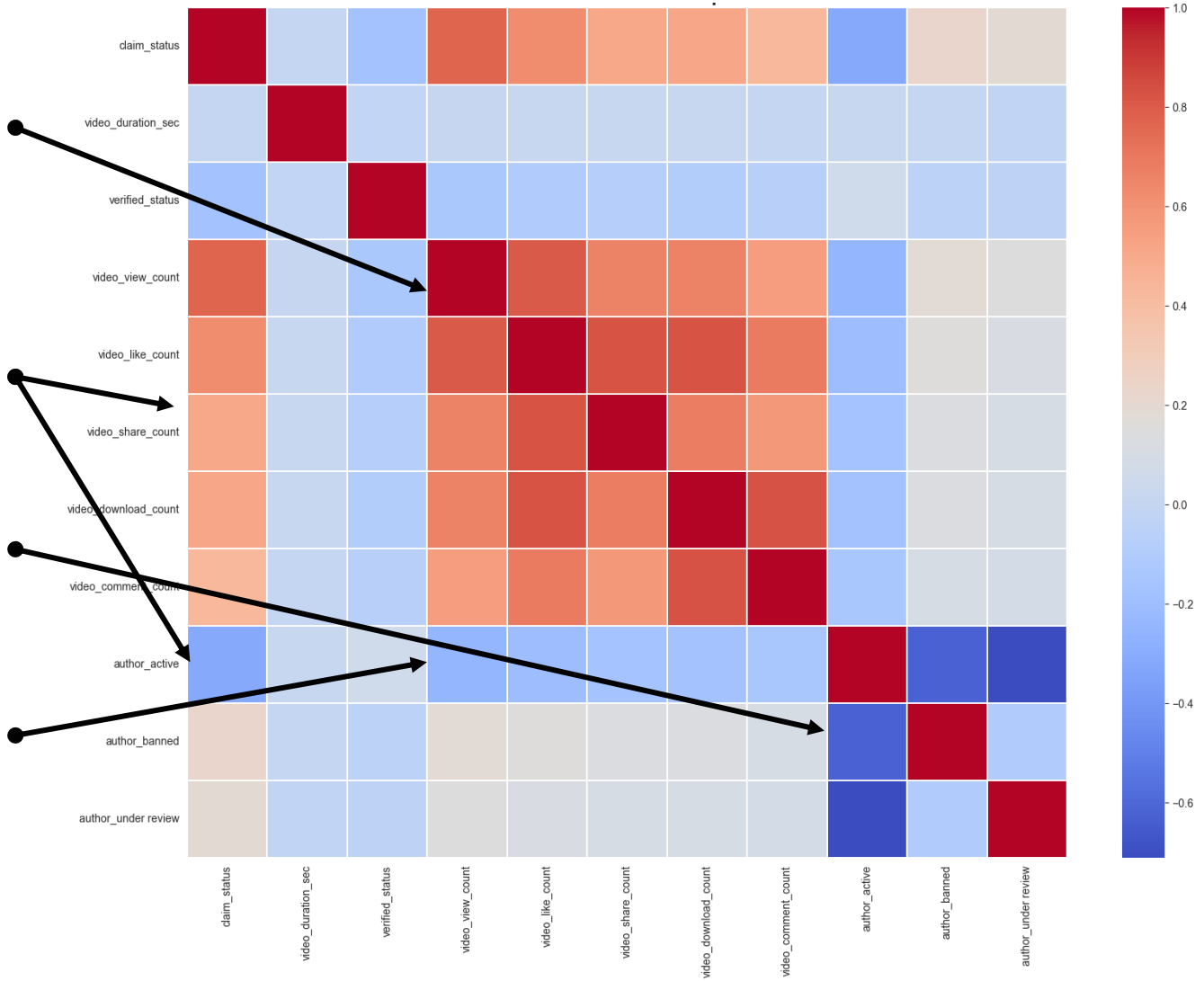
## EDA: Correlation Heatmap

Features like ***video views, likes, downloads, shares,*** and ***comments*** have positive correlations to each other, meaning multicollinearity exist between these **engagement metrics**.

***Claim*** is positive correlated to **engagement metrics** and negative correlated to ***active author***.

***Author status*** features are negatively correlated to each other is predictable.

***Active author*** has slightly negative correlated to engagement metrics.



# Q1. Classify Claim

## EDA: Correlation to Claim Status

| Feature | Correlation to Claim |
|---|---|
| claim status | 1 |
| video view count | 0.76817 |
| video like count | 0.619399 |
| video download count | 0.513217 |
| video share count | 0.512067 |
| video comment count | 0.430487 |
| author banned | 0.230605 |
| author under review | 0.189853 |
| video duration sec | 0.003914 |
| verified status | -0.1706 |
| author active | -0.312438 |

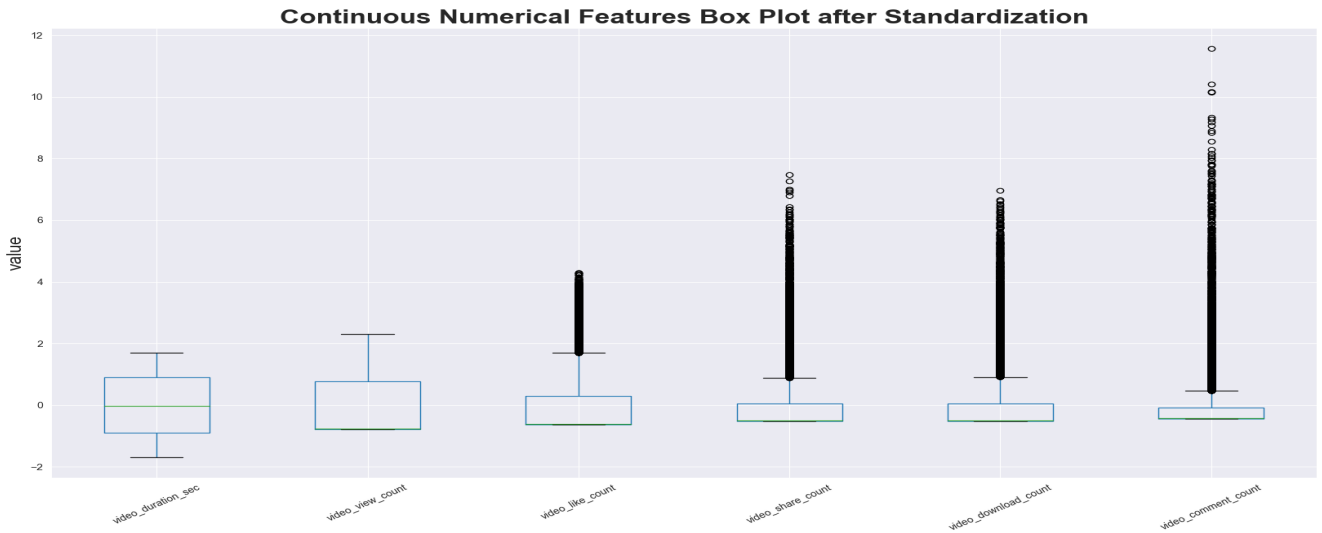
- **Engagement metrics** show strong positive correlation. Claims tend to have higher engagement.
- No real correlation, suggesting **video duration** doesn't significantly affect whether a video is a claim.
- A moderate negative correlation, indicating that claims are less often made by **active authors**.

# Q1. Classify Claim

## Standardization

After standardization:

| claim status | video duration sec | verified status | video view count | video like count | video share count | video download count | video comment count | author active | author banned | author under review |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.637872 | 0 | 0.274362 | -0.486292 | -0.514877 | -0.523104 | -0.436849 | 0 | 0 | 1 |
| 1 | -0.026119 | 0 | -0.352545 | -0.05209 | 0.071757 | 0.055667 | 0.41856 | 1 | 0 | 0 |
| 1 | -0.087748 | 0 | 2.005286 | 0.100327 | -0.433186 | -0.107985 | -0.025402 | 1 | 0 | 0 |
| 1 | -0.457524 | 0 | 0.566138 | 1.166638 | 0.564275 | 0.09209 | 0.2935 | 1 | 0 | 0 |
| 1 | -0.8273 | 0 | -0.614899 | -0.36965 | -0.394104 | -0.250682 | -0.246758 | 1 | 0 | 0 |

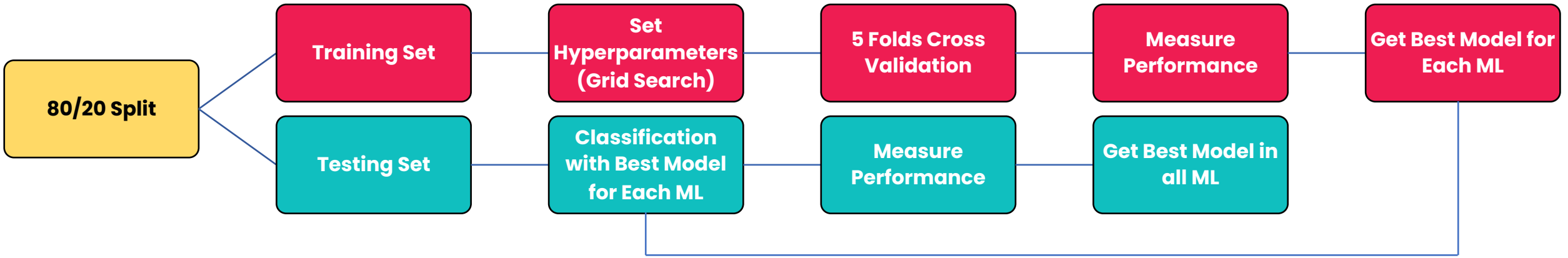



Data Preprocessing → EDA → Standardization → Classification Modeling → Result

# Q1. Classify Claim

## Training

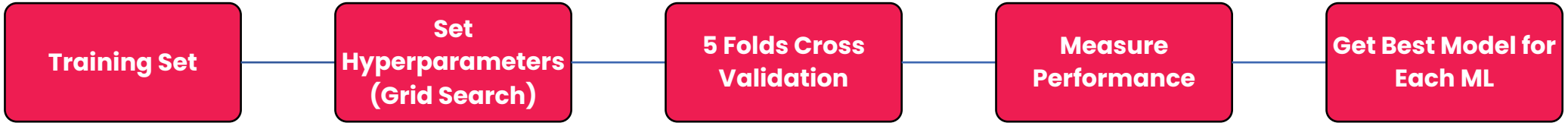
The dataset was divided into independent variables and the target variable ***claim status***. Furthermore, the dataset was split into a training set and a test set, ensuring that the models were trained and validated on different sets of data to avoid overfitting.




# Q1. Classify Claim

## Training

- Hyperparameters for specific models were fine-tuned using a grid search approach.
- Cross-validation with a 5 folds approach was used to prevent overfitting and to ensure that the models generalized well to new, unseen data.

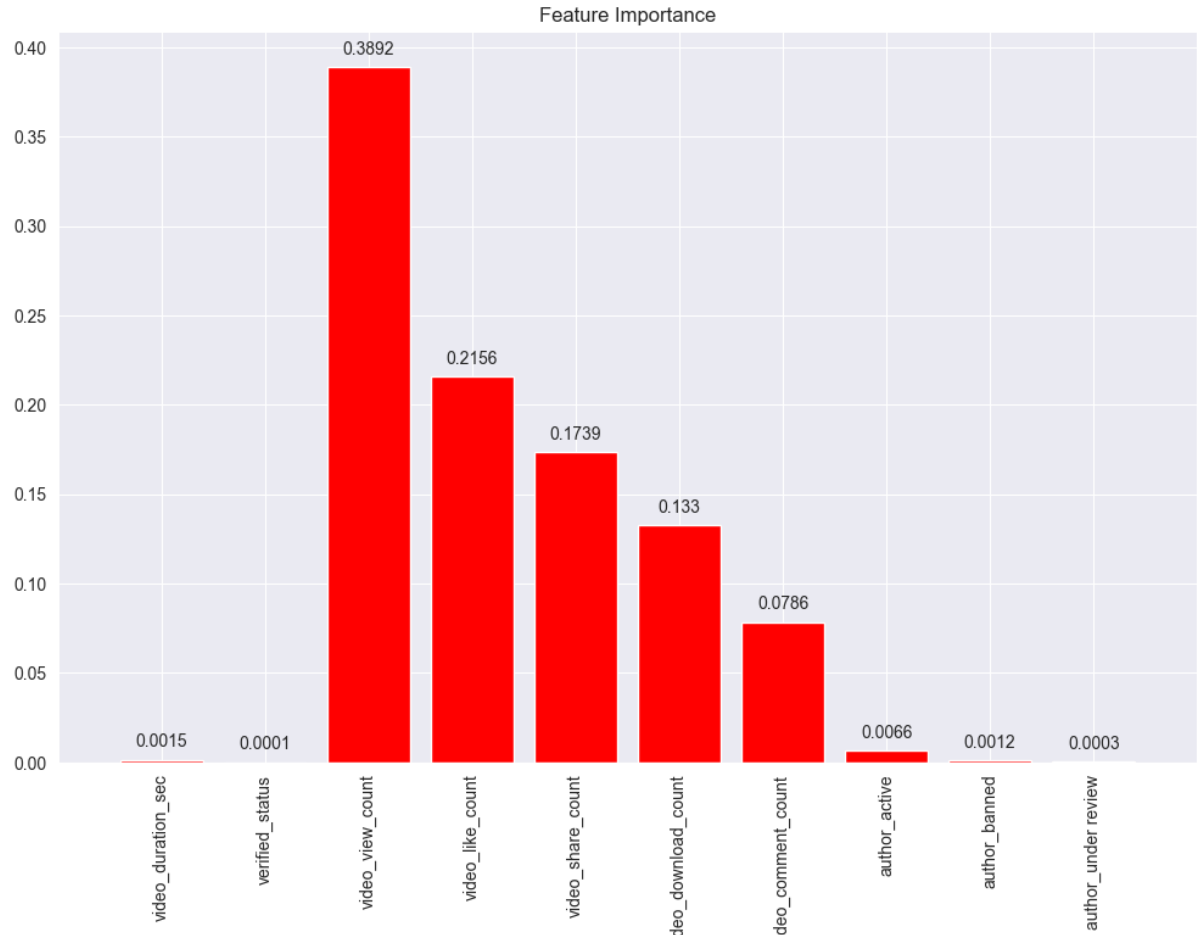



| Model | Search Hyperparameters | Best Hyperparameters | Best CV Accuracy Score |
|---|---|---|---|
| **Random Forest** | n estimators, max depth, min samples split, min samples leaf | n estimators = 200; max depth = None; min samples split = 2; min samples leaf = 2. | **0.9953** |
| **Gradient Boosting** | n estimators, learning rate, min samples split, min samples leaf | n estimators = 200; learning rate = 0.05; min samples split = 2, min samples leaf = 8. | **0.9955** |
| **KNN** | n neighbors, weights, distance type | neighbors = 3; weights= uniform; p = Euclidean distance. | **0.9852** |
| **LDA** | N/A | N/A. | **0.8814** |
| **Logistic Regression** | C, solver | C =100; solver = newton-cg. | **0.9915** |
| **Decision Tree** | max depth, min samples split, min samples leaf | max depth = 10; min samples split = 1; min samples leaf = 2. | **0.9950** |

# Q1. Classify Claim

## Feature Importance after RF

- The number of views a video has is the strongest predictor of the claim status in the model.
- Engagement metrices are all better predictors.
- Video duration, author status,  and verified status are not significant predictors.




Data Preprocessing → EDA → Standardization → Classification Modeling → Result

# Q1. Classify Claim

## Testing

Testing Set → Classification with Best Model for Each ML → Measure Performance → Get Best Model in all ML

| Model | Accuracy CV | Accuracy Train | Accuracy Test | Precision | Recall | F1 Score | AUC |
|---|---|---|---|---|---|---|---|
| Random Forest | 0.9953 | 0.996 | **0.9953** | 1 | 0.9907 | 0.9953 | 0.9982 |
| Gradient Boosting | 0.9955 | 0.9963 | **0.9955** | 0.9995 | 0.9917 | 0.9956 | 0.9984 |
| K-Nearest Neighbors | 0.9852 | 0.9891 | 0.9882 | 0.9995 | 0.9772 | 0.9882 | 0.9914 |
| LDA | 0.8814 | 0.8814 | 0.8866 | 1 | 0.7754 | 0.8735 | 0.988 |
| Logistic Regression | 0.9915 | 0.9916 | 0.9927 | 1 | 0.9855 | 0.9927 | 0.9975 |
| Decision Tree | 0.995 | 0.9967 | 0.9935 | 0.9958 | 0.9912 | 0.9935 | 0.993 |

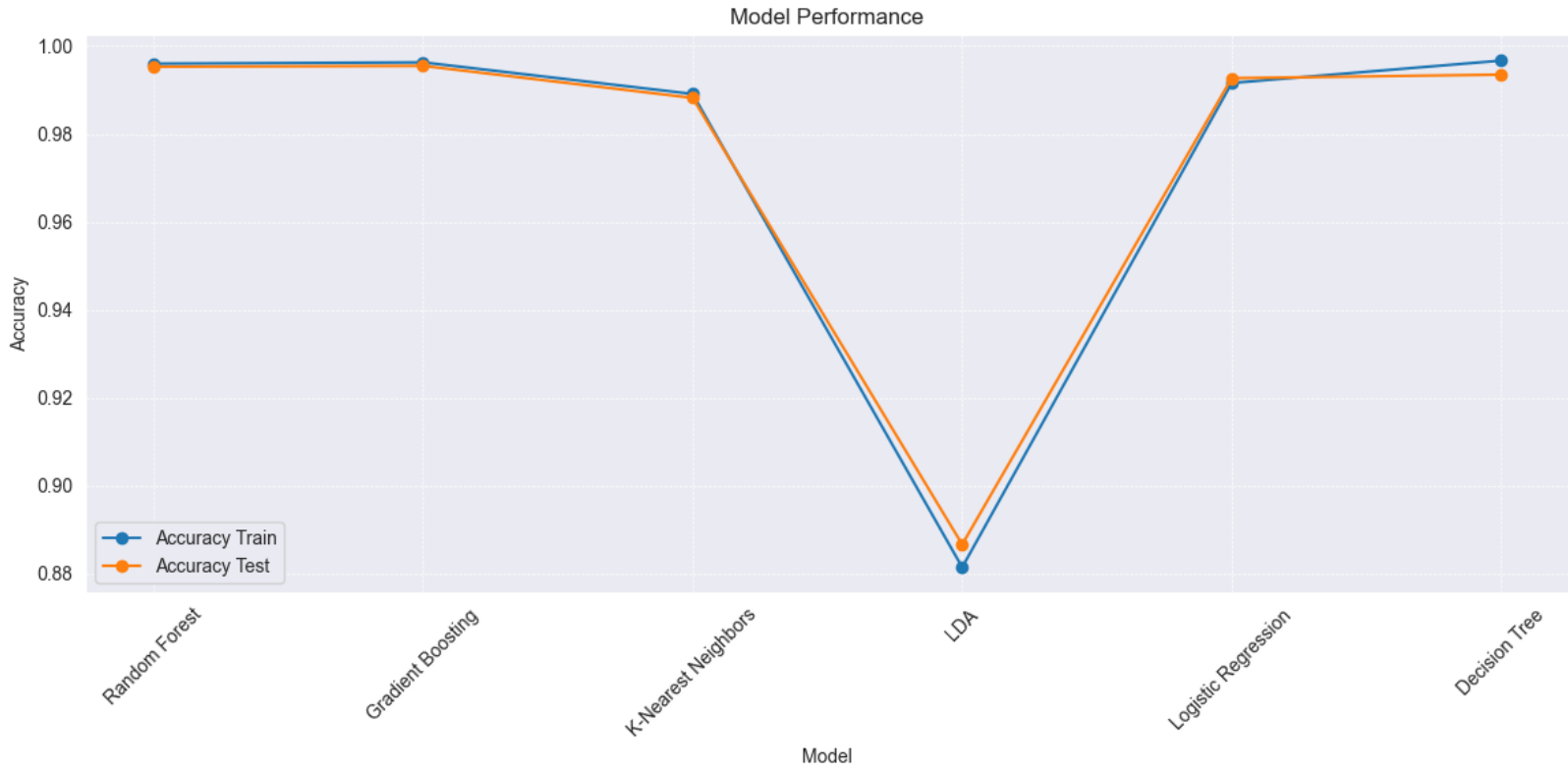
Data Preprocessing → EDA → Standardization → Classification Modeling → Result

# Q1. Classify Claim

## Conclusion

- **Random Forest** and **Gradient Boosting** stand out as the best models. They have high precision, recall, and F1 scores, meaning they balance the rate of true positive identifications with the false positives and false negatives well. Their AUC scores are almost perfect, indicating an excellent ability to discriminate between the classes.

- **Decision Tree** also shows high performance, with the highest training accuracy and very high-test accuracy. It also has a high F1 score and AUC, but it is slightly lower than the Random Forest and Gradient Boosting models, which could indicate a bit more overfitting to the training data.




- **LDA (Linear Discriminant Analysis)** is the weakest model in this set. It has the lowest accuracy scores across training and testing, and its precision is perfect only because it likely predicts one class very well but fails to predict the other

# Analysis and Results
# Q2 Clustering Metrics

# Q2. Clustering Metrics

## EDA: Numerical Data

- **Feature Selection**

  video views, likes, shares, downloads, comments.

- **Standardized Data**

Note: **Data preprocessing, EDA, Standardization** have been conducted in the previous part.

| video view count | video like count | video share count | video download count | video comment count |
|---|---|---|---|---|
| 0.274362 | -0.486292 | -0.514877 | -0.523104 | -0.436849 |
| -0.352545 | -0.05209 | 0.071757 | 0.055667 | 0.41856 |
| 2.005286 | 0.100327 | -0.433186 | -0.107985 | -0.025402 |
| 0.566138 | 1.166638 | 0.564275 | 0.09209 | 0.2935 |
| -0.614899 | -0.36965 | -0.394104 | -0.250682 | -0.246758 |

- **Clustering Process**


**Hyperparameter Setting**

- **Number of Clusters:** From 1 to 10
- **Distance Type:** Manhattan / Euclidean

**Measure Metrics**

- Within-Cluster Sum of Square(WCSS)
- Time Complexity

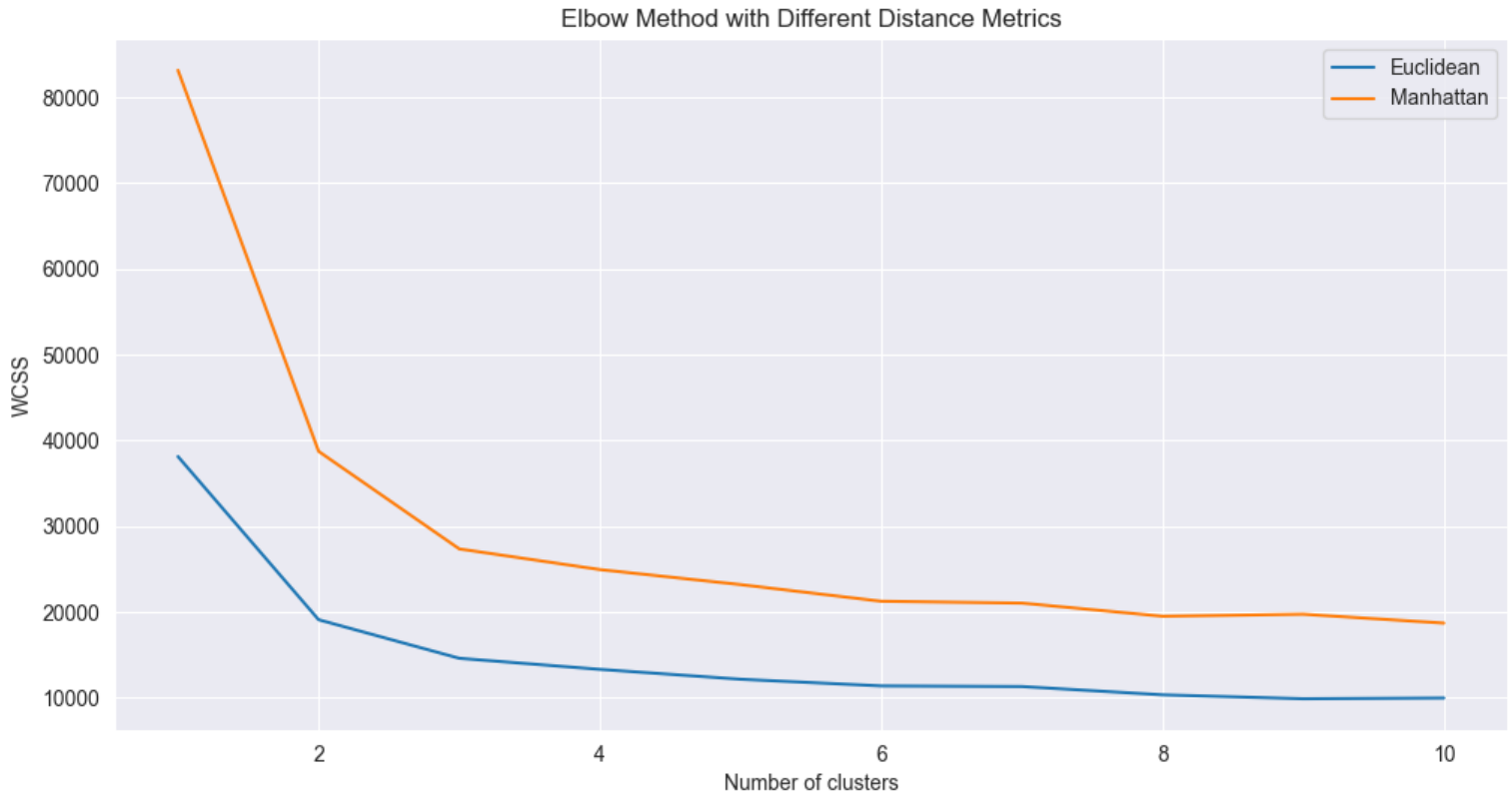
**Analyze Cluster**

- Cluster Centroid Analysis
- Cross Table with Categorical Data
- Pair Plots with Numerical Data

# Q2. Clustering Metrics

## Elbow Method for WCSS

- The **Euclidean distance** line sharply decreases until around **3** clusters, after which the rate of decrease slows down, suggesting that beyond **3** clusters, additional clusters do not significantly improve the model.
- The **Manhattan distance** line shows a similar pattern but with a less pronounced elbow, which could suggest that using Euclidean distance for this clustering problem may yield more distinct and well-separated clusters.

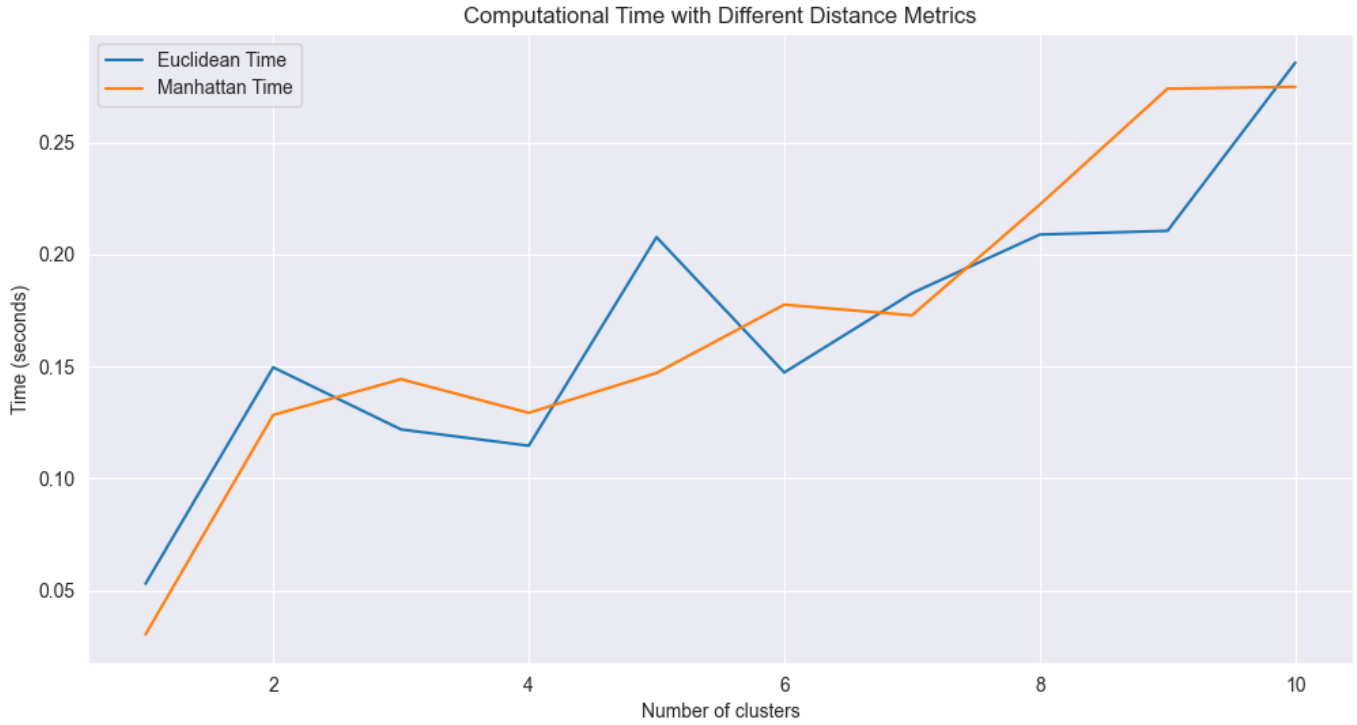



Data Preprocessing → EDA → Standardization → Cluster Modeling → Result

# Q2. Clustering Metrics

## Time Complexity

Both lines increase as the number of clusters increases, which is expected since more clusters generally require more computation for assignment and updating the centroids.




# Q2. Clustering Metrics

## Cluster Centroid Analysis

| Cluster | video view count | video like count | video share count | video download count | video comment count |
|---|---|---|---|---|---|
| 0 | 0.8476 | 0.2567 | 0.1279 | 0.0718 | 0.0121 |
| 1 | -0.7095 | -0.5887 | -0.4891 | -0.4911 | -0.4125 |
| 2 | 1.4981 | 1.9474 | 1.7526 | 1.8492 | 1.6291 |

- **Cluster 0 (C0):** This cluster seems to represent a group of videos with slightly above-average performances. This could be interpreted as moderately popular content.
- **Cluster 1 (C1):** Has negative values for all features, suggesting that videos in this cluster have below-average performance across all engagement metrics. These are possibly less popular or new videos that have not yet gained much traction.
- **Cluster 2 (C2):** Shows significantly higher values across all features, indicating that videos in this cluster are high performers in terms of views, likes, shares, downloads, and comments. This cluster likely represents the most popular and engaging content.

# Q2. Clustering Metrics

## Pair Plots for Numerical Data

- **Cluster 0 (Turquoise)**
  - Has below-average values for all features, as indicated by the centroids and visual representation.
  - Is the most populous cluster, with a high density of points near the origin, suggesting many videos with lower engagement metrics.
- **Cluster 1 (Yellow)**
  - Cluster 0 (Turquoise) Has below-average values for all features, as indicated by the centroids and visual representation. Is the most populous cluster, with a high density of points near the origin, suggesting many videos with lower engagement metrics.
- **Cluster 2 (Red)**
  - Contains videos with the highest engagement metrics across all features. Points are more spread out, indicating a wider range of high-performing videos.
- **Conclusion**
  - This helps in understanding the relationship between different user engagement metrics and can guide targeted strategies for each cluster. For example, strategies for videos in Cluster 0 might focus on initial engagement boosts, while strategies for Cluster 2 might focus on sustaining and leveraging high engagement.

# Analysis and Results
# Q3 Topic Shareability

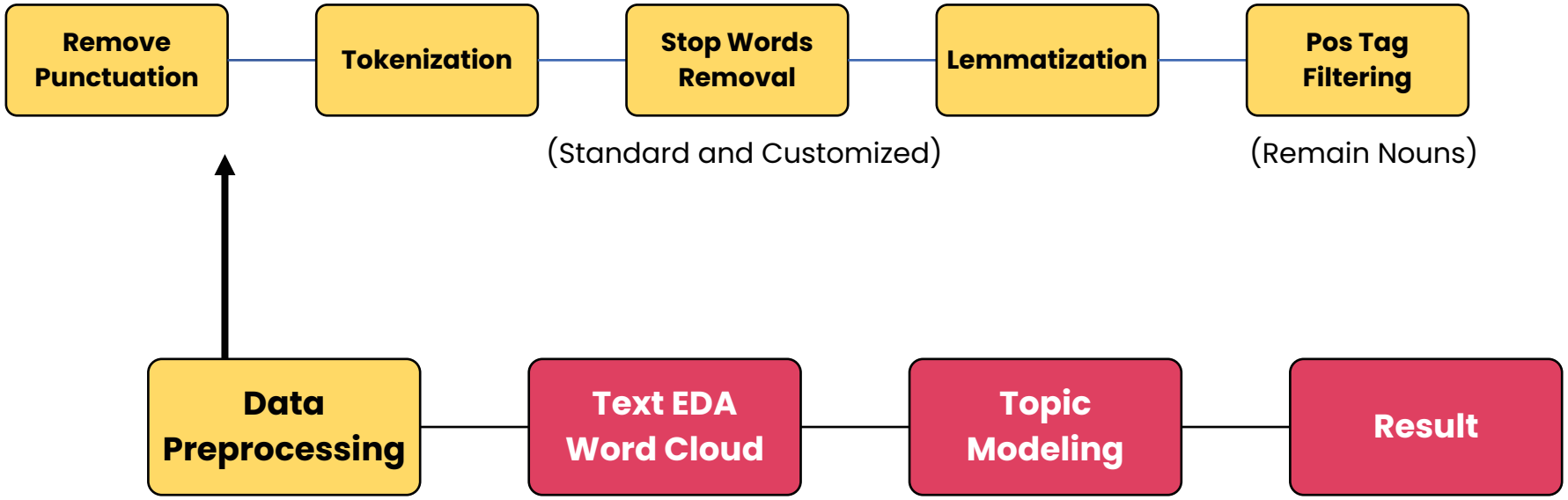
# Q3. Topic Shareability

## Data Preprocessing

- **Feature Selection**
  - **Response variable:** video shares.
  - **Independent variables:** video transcription text.
  - **EDA variables:** claim status, verified status, author status. (Only for Word Cloud analysis)

- **Data Pre-processing**

| video transcription text | video share count |
|---|---|
| someone shared with me that drone ... | 241 |
| someone shared with me that there are ... | 19034 |
| someone shared with me that American industrialist..... | 2858 |
| someone shared with me that the metro ... | 34812 |
| someone shared with me that the number ... | 4110 |

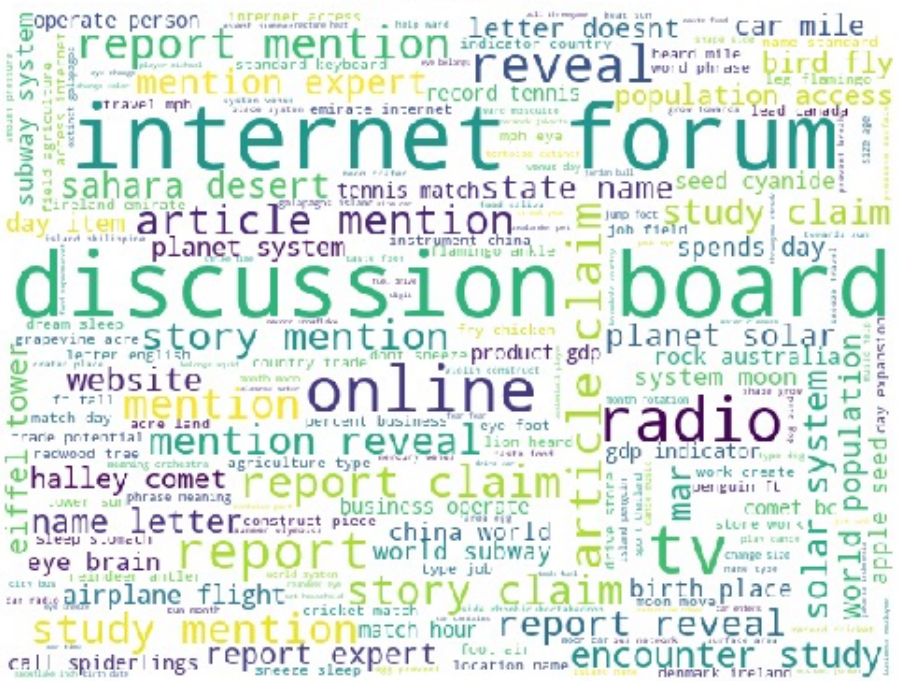


# Q3. Topic Shareability

## EDA: Word Cloud

Opinion Shorts


Claim Shorts

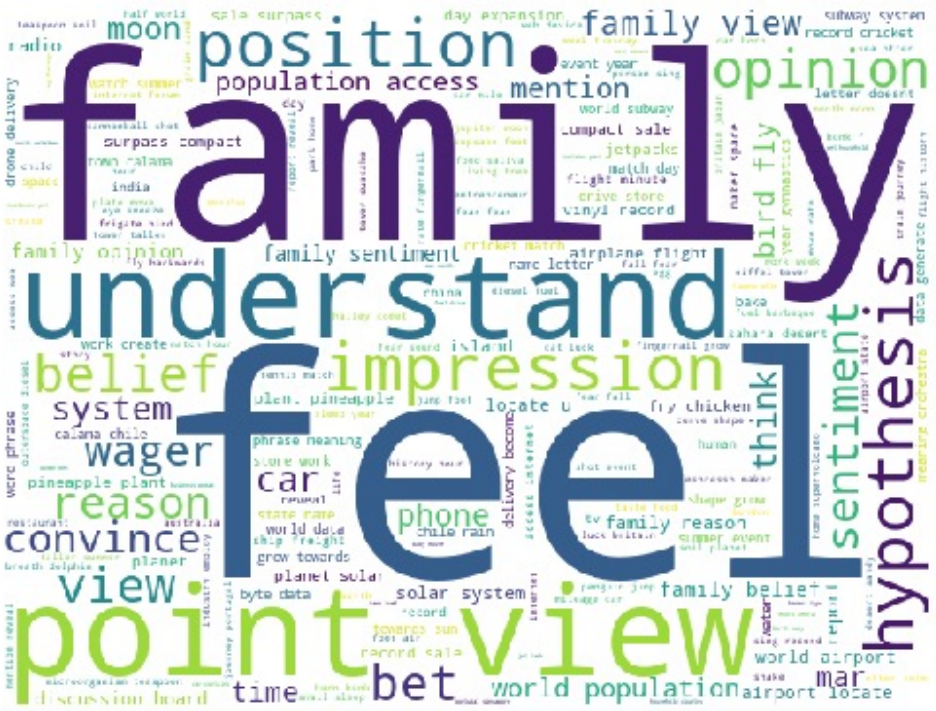



**Data Preprocessing** — **Text EDA Word Cloud** — **Topic Modeling** — **Result**

# Q3. Topic Shareability

## EDA: Word Cloud

Verified Author's Videos



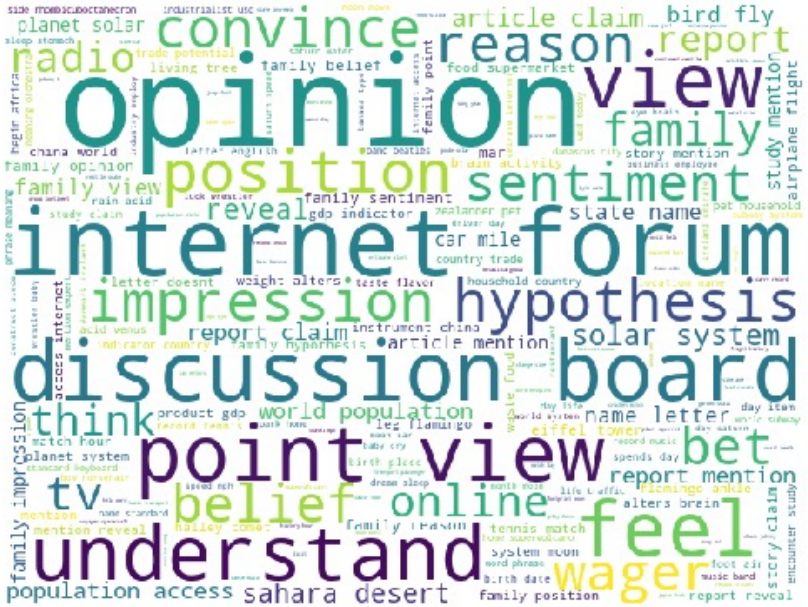

Non-verified Author's Videos


Data Preprocessing → Text EDA Word Cloud → Topic Modeling → Result

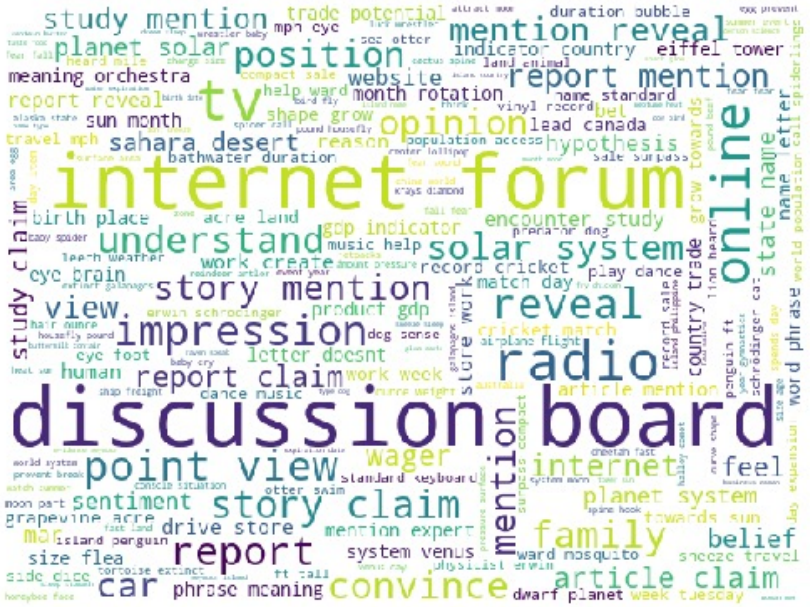
# Q3. Topic Shareability

## EDA: Word Cloud

Active Author's Videos




Under Review Author's Videos




Banned Author's Videos


**Data Preprocessing** — **Text EDA Word Cloud** — **Topic Modeling** — **Result**

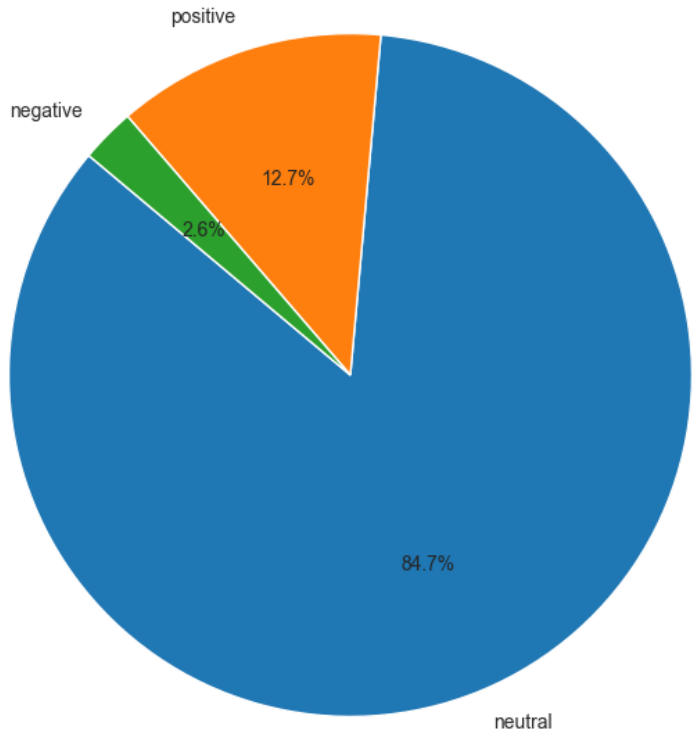
# Q3. Topic Shareability

## EDA: Sentiment Analysis

### Correlation Matrix

| | sentiment | video shares |
|---|---|---|
| **sentiment** | 1 | -0.045 |
| **video shares** | -0.045 | 1 |

### Pie Chart

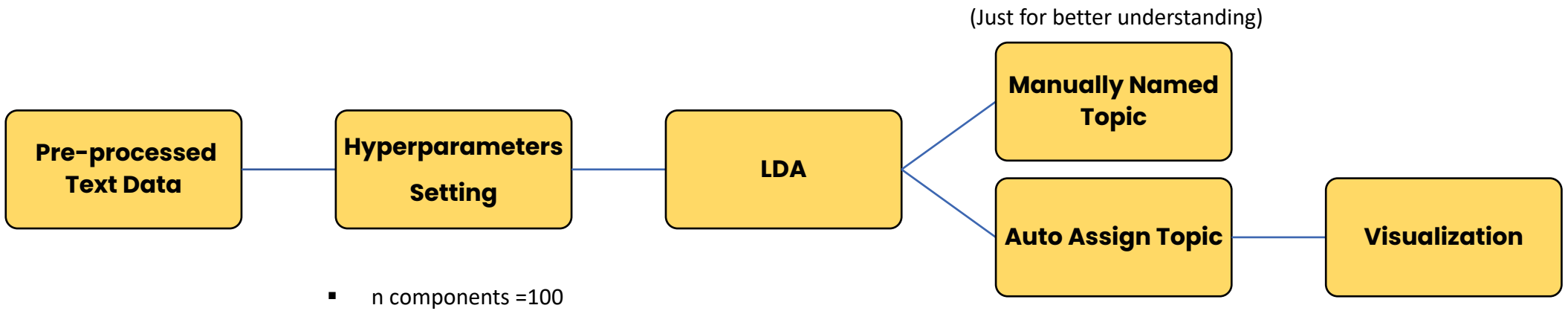



- The correlation coefficient between sentiments and video shares is -0.044933. This indicates a very weak inverse relationship between the sentiment score and the count of video shares.
- This distribution suggests that the texts processed are predominantly neutral, with a smaller but significant portion of positive sentiments and only a few negative ones.

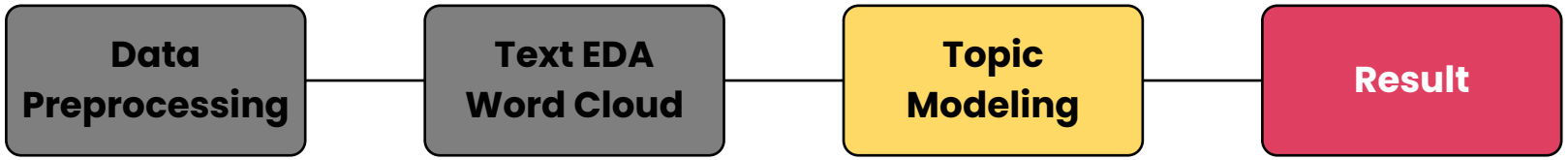
**Data Preprocessing** — **Text EDA Word Cloud** — **Topic Modeling** — **Result**

# Q3. Topic Shareability

## Latent Dirichlet Allocation



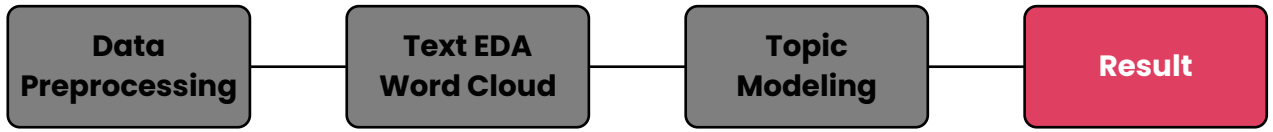

- n components =100
- learning decay = 0.8
- max iteration = 20
- random state = 321



# Q3. Topic Shareability

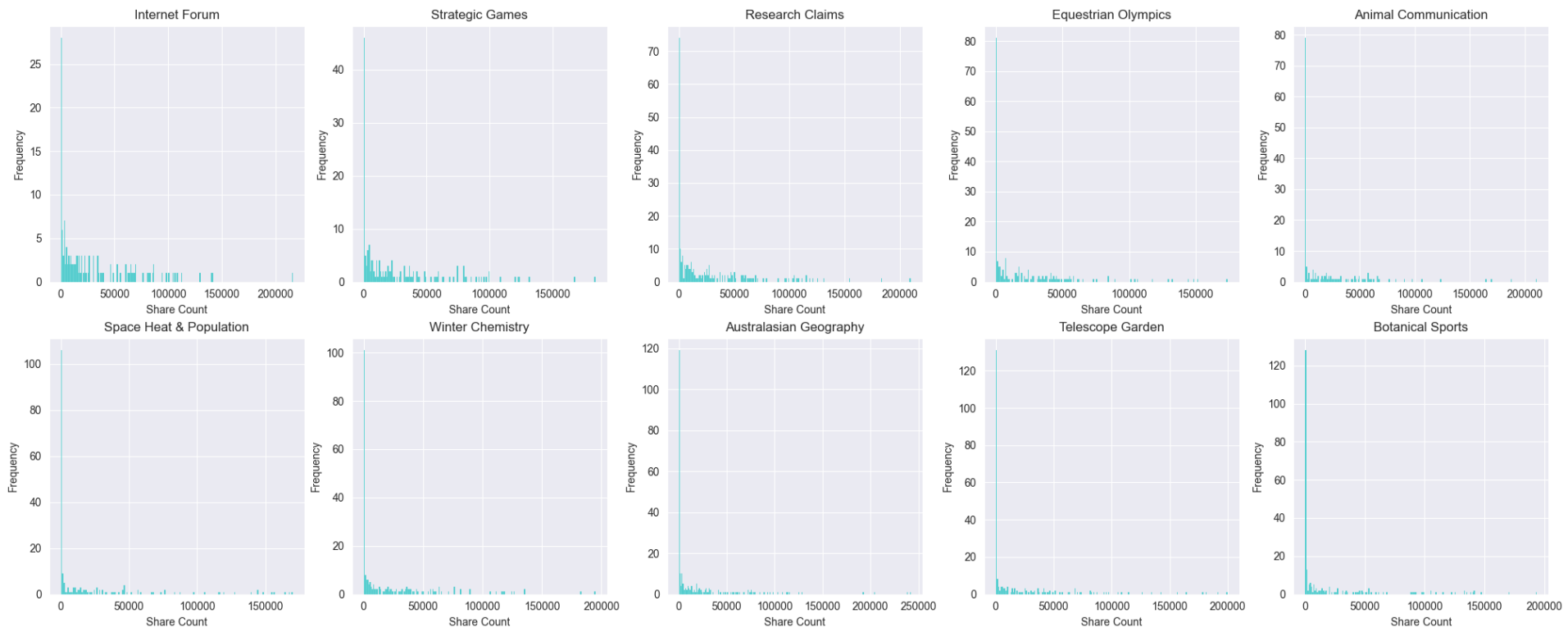
## Top 10 Topic Ranks By Average Shares

| rank | Topic label | Mean shares | Num videos | Std shares |
|---|---|---|---|---|
| 1 | Internet Forum | 30241.51079 | 139 | 37919.1809 |
| 2 | Strategic Games | 27043.56548 | 168 | 35880.83643 |
| 3 | Research Claims | 25488.40928 | 237 | 36327.25757 |
| 4 | Equestrian Olympics | 21119.88442 | 199 | 32993.60845 |
| 5 | Animal Communication | 20900.54545 | 165 | 35906.4816 |
| 6 | Space Heat & Population | 20336.335 | 200 | 38854.1046 |
| 7 | Winter Chemistry | 19861.58108 | 222 | 34054.71367 |
| 8 | Australasian Geography | 19822.168 | 250 | 38023.23944 |
| 9 | Telescope Garden | 19536.64435 | 239 | 38049.11764 |
| 10 | Botanical Sports | 19528.94672 | 244 | 37662.6797 |

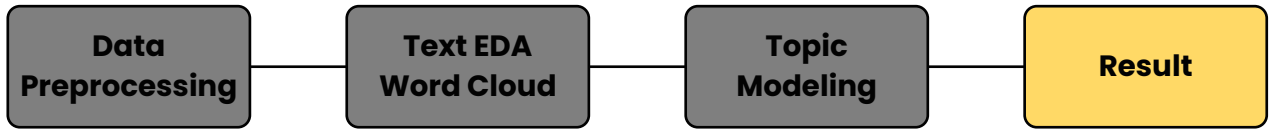


# Q3. Topic Shareability

## Top 10 Topics Histogram

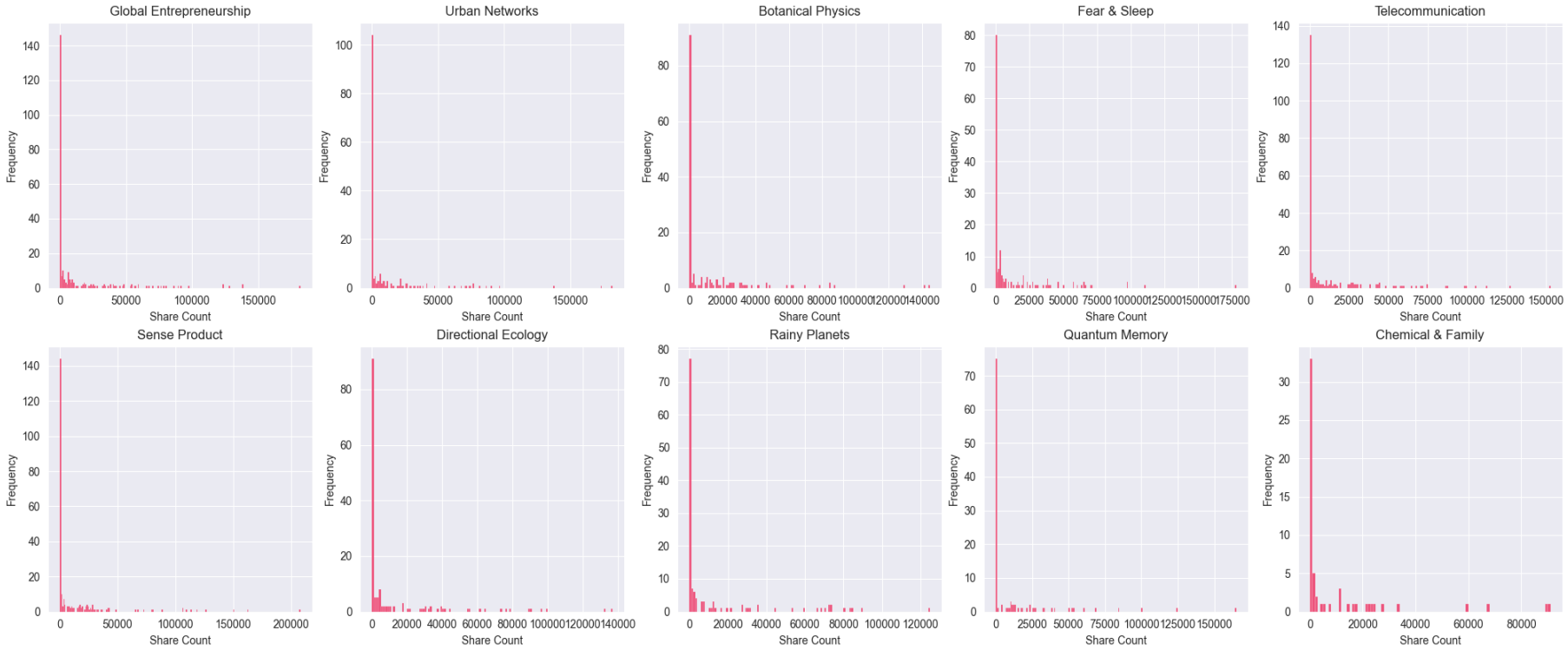



- Topics like "Internet Forum" and "Strategic Games" have a few instances with very high share counts, suggesting that while most videos within these topics are not widely shared, a few have become quite popular.
- The spread of share counts is likely to be wider in the top 10 topics, with more videos achieving a high number of shares.

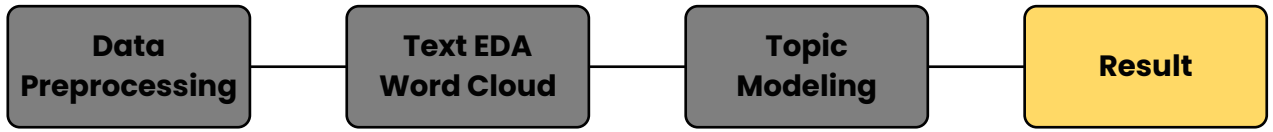


# Q3. Topic Shareability

## Last 10 Topics Histogram




- The top 10 topics such as Internet, Games, Research may contain videos that have the potential to go viral or have already gone viral, as opposed to the last 10 topics such as Planets, Quantum, Chemical  which seem to have a more modest performance in terms of shares.



# Conclusion

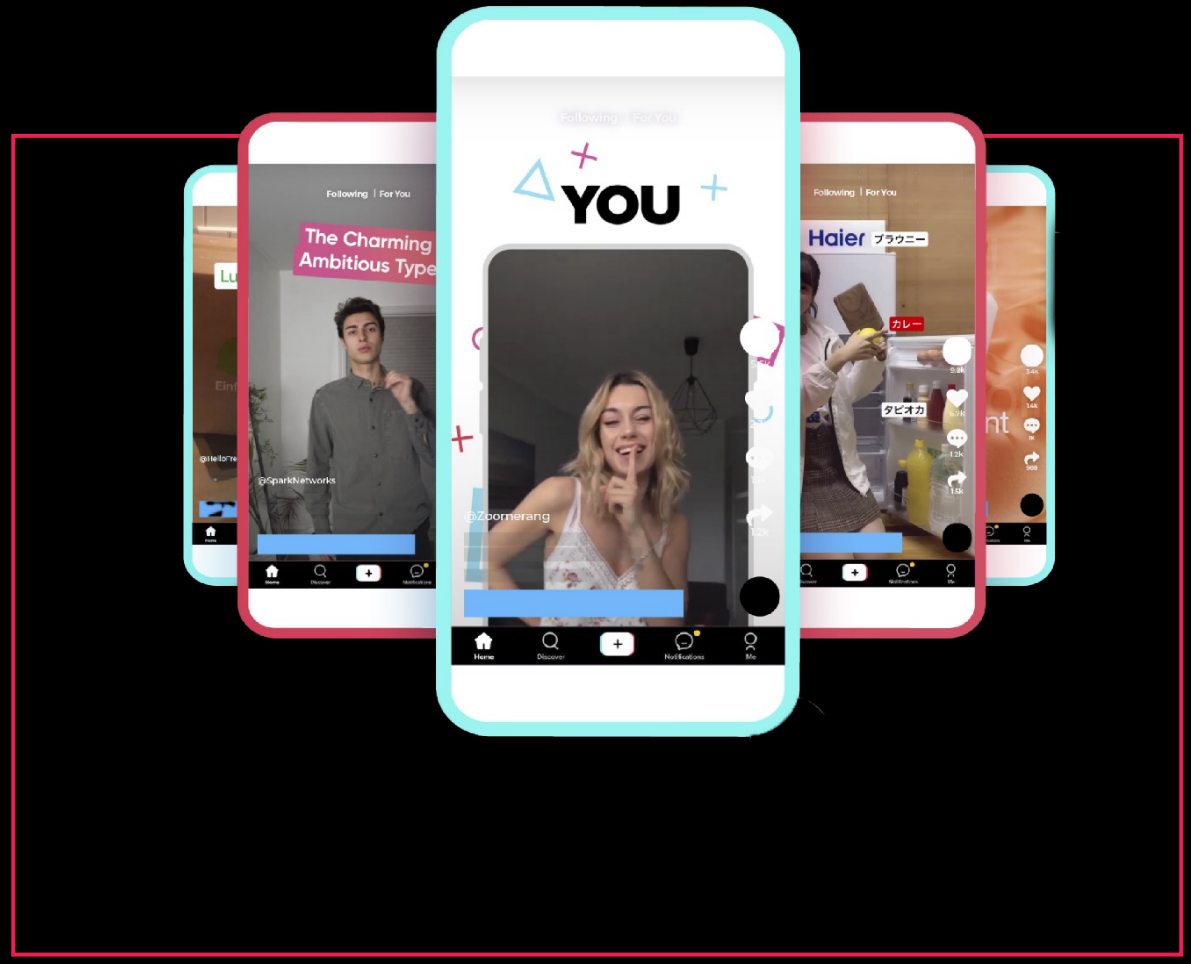
# Conclusion

1. We used tremendous EDA to more understand TikTok video trend.

2. We have properly use Ensemble Learning Model such as Random Forest and Gradient Boosting to classify claim and opinion with 99% accuracy.

3. The final 3 means clustering helps in understanding the relationship between different user engagement metrics and can guide targeted strategies for each cluster.

4. The top 10 topics may contain videos that have the potential to go viral or have already gone viral, as opposed to the last 10 topics which seem to have a more modest performance in terms of shares.



# THANKS