

Comment dataset features:

1. **comment_length** - Longer comments may indicate more thoughtful or engaged responses.
2. **comment_char_length** - The total number of characters in a comment, which shows the level of detail in a comment.
3. **hashtag_count** - Counts the number of hashtags in a comment. Hashtags often indicate participation in trends and can help measure a comment's relevance to a larger discussion.
4. **mention_count** - Counts how many times other users are mentioned in a comment. More mentions might indicate a conversation or reply to others.
5. **sentiment_score_textblob** - Measures the overall sentiment of a comment, from negative (-1) to positive (+1). Helps identify the emotional tone of a comment.
6. **sentiment_score_vader** - Specifically for short texts like social media comments, with a range from -1 (negative) to +1 (positive). Works with emoji, informal words.
7. **positive_reaction** - Base on sentiment_score
8. **negative_reaction** - Base on sentiment_score
9. **neutral_reaction** - Base on sentiment_score
10. **hour_of_day** - Indicates the time the comment was posted. This helps identify patterns in engagement based on the time of day.
11. **day_of_week** - Shows which day of the week the comment was posted. Engagement may vary based on the day, such as higher activity on weekends.
12. **is_weekend** - Marks whether the comment was posted on a weekend (Saturday or Sunday). This can indicate if users are more active on weekends.
13. **is_reply** - Marks whether a comment is a reply to another comment. This helps distinguish between standalone comments and part of a conversation.
14. **lexical_richness** - Measures the diversity of vocabulary used in a comment. Higher values indicate more varied language, suggesting more sophisticated or expressive comments.
15. **emoji_count** - Counts the number of emojis used in a comment.

Video dataset features:

1. **publish_year**: Year in which the video was published. Important for analyzing trends over time.
2. **publish_month**: Month of publication. Useful for identifying seasonal content patterns.
3. **publish_dayofweek**: Day of the week when the video was published. Helps to understand which days might be most effective for posting.
4. **publish_hour**: Hour of the day when the video was published. Can identify optimal publishing times.
5. **is_weekend**: Marks whether the comment was posted on a weekend (Saturday or Sunday). This can indicate if users are more active on weekends.
6. **duration_seconds**: Total duration of the video in seconds. Affects viewer engagement—longer videos may have lower retention.
7. **engagement_rate**: (likes + comments + shares) / views
8. **topicCategories_clean**: Cleaned list of categories assigned to the video. Helps in categorizing content for targeted analysis and recommendations.
9. **video_duration_minutes**: Duration of the video in minutes. Helps to compare engagement levels with different video lengths.
10. **hashtags_in_title**: Number of hashtags present in the video's title. Shows how relevant the video is to hashtag-based search and trends.
11. **hashtags_in_description**: Number of hashtags in the video's description. Important for tracking how well hashtags are used in descriptions.
12. **total_hashtags_tags**: Total number of hashtags from the tags column. Reflects the usage of hashtags in metadata, influencing searchability.
13. **total_hashtags**: Sum of hashtags in the title, description, and tags.
14. **total_mentions**: Number of mentions (user handles) in the description. Indicates the extent of community engagement.
15. **likes_per_view**: Ratio of likes to views. Measures the popularity of the video in relation to its reach.
16. **comments_per_view**: Ratio of comments to views. Shows how much discussion the video generates compared to its reach.
17. **favourites_per_view**: Ratio of favourites to views. Indicates how many viewers find the video particularly valuable.

18. **title_sentiment_score_textblob**: Sentiment score of the video title (TextBlob). Shows if the title has a positive, negative, or neutral sentiment.
19. **description_sentiment_score_textblob**: Sentiment score of the video description (TextBlob). Helps gauge the sentiment of the video's messaging.
20. **title_sentiment_score_vader**: Sentiment score of the title (VADER). Another measure of the title's emotional tone.
21. **description_sentiment_score_vader**: Sentiment score of the description (VADER). Helps analyze the emotional tone of the video description.
22. **popularity_score**: $(\text{views} + \text{likes} + \text{favourites} + \text{comments}) / 4$. Indicates the overall popularity of the video.
23. **topic_categories_count**: Number of topic categories assigned to the video. Helps to understand how diverse or focused the content is.