**1. Introduction**

This project aims to explore the predictive potential of textual and structural features of TikTok news content. Specifically, the goal was to examine whether engagement performance—measured as the log-standardized number of video views—can be explained by content-related variables such as the general topic of the post. The broader motivation of this work is to understand how certain thematic or stylistic aspects of TikTok videos from professional news outlets might contribute to their visibility and reach on the platform.

To approach this question, a multi-stage analytical process was designed. First, the dataset was cleaned and preprocessed to prepare numerical, categorical, and textual variables for modeling. Then, a series of machine learning models were tested, including linear regression, tree-based methods, and feature importance analyses using SHAP values. Finally, a deep learning experiment was conducted using a recurrent neural network (LSTM) to test whether textual information alone could predict standardized view counts.

Although the models did not yield strong predictive performance, the results provide valuable insights into the data structure and limitations of content-based prediction on social media engagement metrics. The following sections document this process and the corresponding findings in detail.

## 2. Dataset Description

The analyses are based on the dataset “News on TikTok: An Annotated Dataset of TikTok Videos from German-Speaking News Outlets in 2023” by Wedel, Mayer, and Batzner (2025), published via GESIS (DOI: [10.7802/2863](https://doi.org/10.7802/2863)). The dataset contains data from over 4000 TikTok videos published in 2023 by leading news outlets from Germany, Austria, and Switzerland. It provides a comprehensive and manually coded collection of metadata, including both descriptive video features (e.g., captions, audio usage, visual style) and theory-driven variables derived from journalism and communication studies (e.g., news values, framing elements, and tone).

Each observation represents one TikTok post, accompanied by a range of variables describing:

* Quantitative metrics, such as number of likes, comments, shares, and views;
* Categorical annotations, including video topic, presence of journalists, and stylistic features;
* Textual fields, such as the general topic description used in this project;
* Temporal information, including upload dates

The dataset was collected under a non-probabilistic, purposive sampling design, covering every second video posted by the most relevant news media accounts within the German-speaking TikTok sphere in 2023. Data collection and annotation were performed by researchers at the Weizenbaum Institute for the Networked Society, with funding from the Federal Ministry of Education and Research of Germany (BMBF). The dataset is publicly available under a CC BY-NC-SA 4.0 license.

**3. Exploratory Data Analysis**

**3.1 Data Cleaning and Preparation**

The dataset originally contained data from 4286 TikTok videos from various news outlets. After removing rows with missing values (308 cases), 3978 valid observations remained for further analysis. Invalid numerical codes (e.g., -99, 77) that indicated missing information were subsequently converted to NaN and dropped as well. The resulting dataset was saved as tiktok\_clean.csv for modeling and further exploration.

**3.2 Descriptive Statistics**

***3.2.1 Categorical Variables***

Several categorical variables were inspected to understand the overall composition of the dataset.

* **News outlet (V3):** The most frequent source was *20minuten* (21.3%), followed by *heute.at* (12.9%), while all others occurred less frequently.
* **News classification (V5):** All items were categorized as “news-related content,” confirming the dataset’s thematic consistency.
* **Visual set-up (V6):** Roughly half of all videos were *journalistic-centric*, while about one-third showed *live footage without journalistic elements.*
* **News format type (V10):** A clear majority (73%) were *newsflashes*.
* **Geographic scope (V12):** Most posts dealt with *domestic* topics (56%), followed by *foreign news* (29.5%).
* **Temporal scope (V18):** The majority (82%) referred to *current events*, while past or future topics were rare.

These distributions suggest that TikTok news content in this dataset primarily focuses on short, journalistic video formats centered on domestic and current topics.

***3.2.2 Binary Variables***

Binary indicators were used to describe the presence or absence of specific visual, audio, and interactive features. A heatmap of mean values across all binary features revealed several key tendencies:

* **Visual features:**
  + Text overlays were almost always present (96%).
  + Screenshots (89%) and infographics/animations (93%) were rare.
  + Channel logos were used in about 90% of all cases.
* **Audio features:**
  + Journalists’ voices were common (83%).
  + Other people’s voices were less frequent (26%).
  + Background music appeared in 70% of the videos, while sound effects and ambient sounds were less common.
* **Interactive features:**
  + Mentions and calls for user interaction were rare (below 25% for all categories).
* **Content features:**
  + Positive and negative framing occurred in only a minority of videos.
  + Conflicts and follow-up stories were almost absent (below 20%).
  + “Scope” indicators (how many people are affected) were present in about 82% of the videos.

Overall, the data suggest a dominant reliance on spoken news presentations supported by text overlays and background music, but with little audience interaction or emotional framing.

***3.2.3 Numeric Variables***

Four numeric variables described engagement metrics: likes, views, comments, and shares.  
Descriptive statistics showed a pronounced right-skewed distribution across all measures, with a few videos receiving disproportionately high engagement.

**Table 1**

*Descriptive Statistics of TikTok Video Performance Variables*

| Variable | Mean | Std. Dev. | Min | Max |
| --- | --- | --- | --- | --- |
| Like count (M4) | 4916.55 | 20747.55 | 1 | 821805 |
| View count (M5) | 51576.22 | 226488.97 | 0 | 7105469 |
| Comment count (M6) | 229.28 | 2271.87 | 0 | 137075 |
| Share count (M7) | 320.61 | 2016.50 | 0 | 66363 |

Note. All metrics are based on raw TikTok engagement counts per video. *N* = 3868.

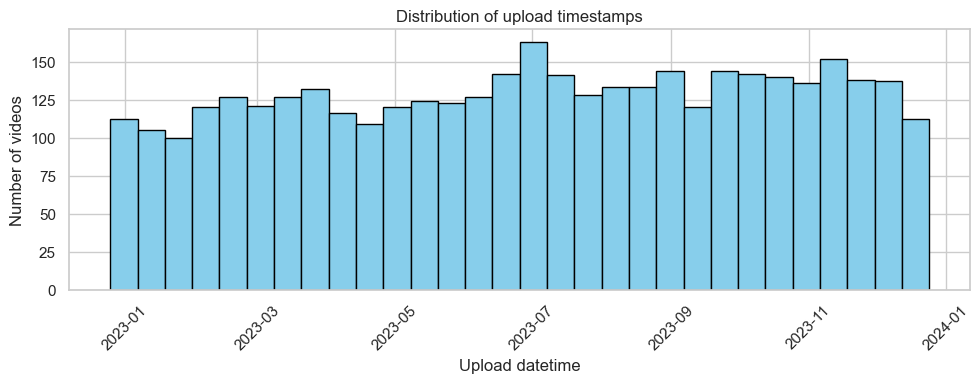
A log-transformation (log1p(x)) was applied to all engagement metrics to reduce skewness and mitigate the impact of outliers. The resulting distributions appeared approximately normal, suggesting the transformation was appropriate for subsequent modeling.

**3.3 Temporal Distribution of Posts**

The posting date variable (M3\_create\_time) was converted to a datetime format, allowing an inspection of posting patterns over time. A monthly distribution revealed that videos were collected over a roughly one-year period between 2022 and 2023, with relatively stable posting activity across months. Peaks in posting frequency corresponded to periods of increased news activity (e.g., major political or social events).

**Figure 1**

*Distribution of TikTok Upload Dates Across 2023.*



**3.4 Textual Variable (“General Topic”)**

The variable V11\_general\_topic contains free-text descriptions or short thematic summaries. The most common keywords referred to interview formats (e.g., “either–or” interviews), while the remainder consisted of diverse and heterogeneous entries that could not easily be grouped into consistent topical categories. This heterogeneity likely limits the predictive potential of textual content in modeling engagement metrics.

**3.5 Dataset Cleaning**

Before modeling, the dataset underwent a data cleaning procedure to ensure consistency and validity.  
Specifically, invalid codes representing ambiguous or residual categories (e.g., *-99 = unclear*, *77 = other*) were systematically replaced with missing values. All records containing missing values after this replacement were excluded from the analysis, resulting in a fully complete dataset with consistent categorical and numerical fields. These excluded cases represented a very small fraction of the dataset and did not contain interpretable or conceptually distinct information, making their removal methodologically sound.

**3.6 Summary**

The exploratory analysis provides several insights into the dataset’s structure:

1. The data are heavily imbalanced toward domestic, current, and short-form news content.
2. Engagement metrics are highly skewed, requiring log transformation.
3. Visual and audio elements follow clear stylistic conventions (text overlays, background music, and presenter voices).
4. The “general topic” variable is diverse and unstandardized, which poses challenges for text-based predictive modeling.

Overall, these results reveal that the dataset captures a very specific subgenre of TikTok news videos—short, text-heavy, domestic news updates—but offers limited variation in features that might drive engagement patterns.

**4. Data Preparation and Feature Engineering**

**4.1 Overview**

This step prepares the cleaned TikTok dataset (tiktok\_clean.csv) for regression analysis.  
The goal is to create a structured, numerical dataset containing only relevant and interpretable predictors while ensuring consistency across categorical, binary, and temporal variables.

**4.2 Data Import and Initial Cleaning**

The cleaned dataset is first loaded into a pandas DataFrame. Two variables that are not relevant for analysis — the coder ID and the video link — are dropped. These contain administrative or external information that does not carry analytical value.

**4.3 News Outlet Grouping**

The variable V3\_news\_outlet contains the outlet names of the news producers.  
Since some outlets occur very infrequently, categories with less than 4% of total observations are grouped into a new “other” category. The grouped variable replaces the original, and dummy variables are created via one-hot encoding. This approach reduces noise from rare categories while preserving meaningful group differences.

**4.4 Country of Origin and News Classification**

The variable V4\_country\_of\_origin is one-hot encoded to create binary indicators for each country.  
The variable V5\_news\_classification is dropped because it contained only a single category after cleaning and thus provides no variance.

**4.5 Visual Set-Up**

The categorical variable V6\_visual\_set-up is one-hot encoded with the prefix “visual”. This preserves interpretability of different visual presentation styles (e.g., anchor-based vs. image-based).

**4.6 Binary Content Features**

Several binary variables already exist in 0/1 format, such as indicators for the presence of images, text overlays, or music. To understand their prevalence, proportions were calculated for each.

**4.7 Audio and Interactivity Features**

Similar summaries were generated for audio-related (V8a–V8e) and interactive (V9a–V9c) variables.  
All are kept as binary predictors since they directly describe content composition and engagement prompts.

**4.8 News Format and Geographic Scope**

Two additional categorical variables — the news format (V10\_news\_format\_type) and the geographic scope (V12\_geographic\_scope) — were also one-hot encoded:

**4.9 Temporal and Contextual Features**

V18\_temporal\_scope is simplified into a binary variable representing whether the post refers to a **current** event. Additionally, the upload date (M3\_create\_time) is transformed into a numeric “days since upload” feature. This variable captures content recency, which can influence exposure and engagement levels.

**4.10 Performance Metrics**

Four performance indicators are available in logarithmic form:

* M4\_like\_count\_log
* M5\_view\_count\_log
* M6\_comment\_count\_log
* M7\_share\_count\_log

A correlation matrix shows strong intercorrelations between them, suggesting they measure similar underlying performance dynamics. To avoid redundancy, the log-transformed **view count** is selected as the central performance measure and standardized to a z-score.

**4.11 Final Feature Selection**

A subset of relevant predictors is compiled for the regression dataset, including:

* Categorical dummies (outlet, country, visual set-up, format, geography)
* Binary indicators (images, music, calls to action, etc.)
* Contextual variables (temporal scope, recency)
* The standardized target (M5\_view\_count\_log\_z)

Boolean columns are converted to integers to ensure full numerical compatibility for regression models.

**5. Modeling Pipeline**

**5.1 Objective**

The purpose of this step was to build predictive models for video performance on TikTok, represented by the standardized log of view counts (M5\_view\_count\_log\_z). Two types of models were used: a **linear regression model** (for interpretability) and a **random forest model** (for predictive performance and non-linear effects).

**5.2 Data Preparation**

The regression-ready dataset from the previous preprocessing step was loaded.  
A standardized version of the variable days\_since\_upload was created (days\_since\_upload\_z) to control for the recency of publication. The raw (non-standardized) version was removed to avoid redundancy. All categorical and binary predictors were already dummy-encoded in the previous step, so the feature matrix X contained only numeric values. The target variable was defined as M5\_view\_count\_log\_z.   
The data were split into a training set (70%) and a test set (30%).

**5.3 Multicollinearity Check**

Variance Inflation Factors (VIFs) were computed to detect collinearity among predictors.  
Some dummy variables—particularly V7b\_text, newsformat\_2.0, and country\_of\_origin\_CH—showed elevated VIF values (> 10), which suggests overlap in information content.  
Nevertheless, all predictors were retained for interpretability, as multicollinearity mainly affects coefficient precision rather than prediction accuracy.

**5.4 Linear Regression Results**

The ordinary least squares model achieved:

* **Train R²:** 0.65
* **Test R²:** 0.65

This means that roughly two-thirds of the variation in standardized view counts could be explained by the predictors. Several variables were statistically significant, as can be seen in Table 2. The adjusted R² = 0.65 indicates a well-fitting yet interpretable model. Residual diagnostics (Durbin-Watson ≈ 2.0) showed no major autocorrelation.

**Table 2**

*Linear Regression Results Predicting Standardized Log View Counts*

| **Variable** | **Coefficient (β)** | **p-value** | **Interpretation** |
| --- | --- | --- | --- |
| days\_since\_upload\_z | +0.69 | < 0.001 | Older videos had accumulated more views |
| V8a\_journalists\_moderators\_voice | +0.15 | < 0.001 | Presence of a journalist’s voice slightly increased performance |
| V8e\_ambient | +0.10 | 0.002 | Background ambience was positively associated |
| geo\_1.0 | +0.14 | < 0.001 | Regional scope 1 correlated with higher reach |
| Most news outlets | −0.5 to −1.0 | < 0.01 | Certain outlets consistently underperformed relative to the reference group |

*Note*. Results from an OLS regression model. Coefficients (β) represent standardized effects on the dependent variable.

**5.5 SHAP Analysis (Linear Model)**

A SHAP analysis was performed to quantify the contribution of each predictor to model predictions.  
The SHAP summary plot visualizes feature importance and direction of influence.  
The most influential variables were days\_since\_upload\_z, the outlet dummies, and several audio features, confirming the regression coefficients’ patterns.

**5.6 Random Forest Baseline**

A baseline random forest regressor (500 trees) was trained to capture possible non-linear relationships.  
Performance metrics:

* **Train R²:** 0.98
* **Test R²:** 0.84

The model generalized well, showing a substantial gain in predictive power compared to the linear model, while maintaining a moderate overfitting gap (0.13 R² difference).

**5.7 Random Forest Hyperparameter Optimization**

A randomized grid search was performed across 30 parameter combinations. The best model configuration was:

n\_estimators = 200

max\_depth = 20

min\_samples\_split = 5

min\_samples\_leaf = 4

max\_features = 0.5

The tuned random forest achieved:

* **Train R²:** 0.90
* **Test R²:** 0.85

Thus, the tuning slightly reduced overfitting while maintaining strong predictive accuracy.

**5.8 SHAP Analysis (Random Forest)**

SHAP values were again computed to interpret the non-linear model.  
days\_since\_upload\_z remained the dominant predictor, followed by outlet categories and certain audio/visual features. This confirms the strong temporal effect and highlights outlet-specific performance differences even when controlling for content variables.

**5.9 Summary of Findings**

* **Temporal recency** is the most decisive factor for TikTok view counts.
* **Outlet identity** systematically influences performance.
* **Audio design** (journalist voice, ambience) matters more than visual style.
* Non-linear models (Random Forest) provide a clear accuracy advantage while maintaining interpretability through SHAP.

**6. Deep Learning Analysis – Text-only LSTM**

**6.1 Objective**

This final analysis tested whether the short free-text field (V11\_general\_topic) alone contains enough semantic signal to predict standardized log view counts (M5\_view\_count\_log\_z). The approach used a small LSTM network trained from scratch on tokenized topic strings.

**6.2 Data and preprocessing**

* Source: tiktok\_clean.csv (same cleaned dataset used in earlier sections).
* Target: M5\_view\_count\_log\_z (z-standardized log view count).
* Text field: V11\_general\_topic (free-text topic descriptions).

Text statistics:

* Maximum tokens per description: **18**
* Mean tokens per description: **5.03**
* Tokenizer vocabulary used: **max\_words = 500**
* Input shape to the model: **(3868, 18)** (3,868 samples padded/truncated to length 18)

Because the text entries are extremely short and the token set is sparse, the vocabulary was intentionally restricted and sequences were padded to the dataset maximum length to avoid trimming potentially informative tokens.

**6.3 Model specification and training**

* Tokenization: Tokenizer(num\_words=500, oov\_token="<OOV>").
* Padding: pad\_sequences(..., maxlen=18, padding='post').
* Model architecture (small LSTM regression):
  + Embedding layer (embedding\_dim = 64)
  + LSTM(64)
  + Dropout(0.2)
  + Dense(1, activation='linear')
* Loss / optimizer: mean squared error (MSE) / adam.
* Regularization / early stopping: EarlyStopping(monitor='val\_loss', patience=4, restore\_best\_weights=True).
* Training regime: up to 30 epochs, batch size 32. Training stopped early (restored best weights at end of epoch 1).

**6.4 Results**

***6.4.1 Training behavior***

* The training loss and MAE decreased across epochs (train MSE dropped toward ~0.79 and train MAE toward ~0.75).
* Validation loss and MAE remained roughly constant (val\_loss ≈ 1.02, val\_mae ≈ 0.85).
* EarlyStopping restored weights from epoch 1 (best validation performance).

***6.4.2 Final test performance***

* **Test R²:** **−0.0002** (≈ 0)
* **Test MSE:** **1.0154**
* **Test MAE:** **0.8510**

Interpretation of metrics:

* An R² value ≈ 0 indicates the model performs effectively no better than predicting the mean of the target distribution; negative values would mean worse than mean, so −0.0002 means essentially no explanatory power.
* Test MSE ≈ 1.015 is close to the variance of the standardized target (which is 1), further confirming that predictions are close to the sample mean rather than capturing meaningful variation.
* Train vs. validation behavior suggests the model fit training idiosyncrasies but did not learn patterns that generalize to held-out data.

**6.5 Diagnostic interpretation**

The average description is ~5 tokens (max 18). LSTMs rely on sequential context; with so few tokens there is limited sequential information to learn. The V11\_general\_topic field is irregular and heterogeneous (many unique short phrases). There is likely little systematic mapping from this brief description to eventual view counts. Moreover, many tokens are infrequent; embeddings trained from scratch cannot form reliable representations with such sparse data. Earlier models showed that variables like days\_since\_upload and outlet identity explain most variance. Text alone is a weak predictor in this dataset. Finally, while 3,868 samples is moderate, for training embeddings and LSTM weights from scratch on short text this is often inadequate.

**6.6 Conclusion**

The LSTM deep-learning attempt correctly demonstrates an important empirical lesson: **not every dataset benefits from a from-scratch neural network**. Short, heterogeneous texts and dominant time/outlet effects mean that simpler or transfer-learning approaches are more promising. The negative result is informative — it shows where predictive effort should be concentrated next (pre-trained embeddings, richer non-text features, or alternative targets).