Deep Learning for Influencer Marketing Analysis Using Global YouTube Trending Data

Course : BA-64061-001- Advanced Machine Learning
Student : Tarun Betageri Tejeswara
Professor : Chaojiang (CJ) Wu, Ph.D.
Date : December 3, 2025

TABLE OF CONTENT

ABSTRACT

This paper presents a multimodal deep learning model for viral potential prediction of YouTube videos based on metadata and text (title, tags, description) along with global engagement statistics. We build four models (MLP, LSTM, CNN, FUSION) using more than six million instances from the YouTube Trending Videos Global Dataset. MLP uses metadata, LSTM uses text, CNN uses text, and FUSION model uses metadata and text. The model fusion achieved an accuracy of 98.26% and AUC score of 0.9967, showing the efficacy of combining structured metadata and unstructured content in influencer marketing analytics.*(Chandrasekaran et al., 2023; Liu et al., 2021).*

EXECUTIVE SUMMARY

Developed an end-to-end deep learning pipeline to predict virality for YouTube videos using one of the largest publicly available influencer-marketing datasets, the Global YouTube Trending Videos dataset (over 6.3 million videos from multiple countries) (*Kumar et al., 2017*). With influencer marketing being a $24+ billion industry, brands need to rely on accurate, data-driven predictors to select which content and creators to work with for a highly engaged audience. This article addresses this need by constructing predictive models that use a combination of structured metadata (views, likes, comments, posted time and category) and unstructured text data (video titles, tags and descriptions) *(Zhang & Zhao, 2020).*

The list included: (1) a metadata-only multi-layer perceptron MLP, (2) a text-based long short-term memory LSTM, (3) a text-based convolutional neural network CNN and (4) a multimodal Fusion Model that combined metadata and text signals. The results found that the Fusion Model outperforms all other models by 98.26% accuracy, area under ROC curve (AUC) of 0.9967 and highest precision and recall *(Bengio et al., 2016; Hochreiter & Schmidhuber, 1997).* These findings suggest that virality is best explained by a combination of behavioral engagement patterns and semantic content information.

The study assists brands, agencies, and online platforms for finding the best content creators and making improved campaign decisions, predicting post-performance, and improving content recommendation systems on social media platforms. This study shows multimodal deep learning may transform influencer marketing analytics. Multimodal deep learning may help with data-driven planned decision-making in a digital landscape. The digital landscape is evolving rapidly.

INTRODUCTION

In influencer marketing, brands rely increasingly on creators for engagement of people on platforms. Platform analytics play an important role for identifying content creators to partner with *(Truong & Simmons, 2010; Liu et al., 2021)*. Predicting content virality is challenging because of the non-linear nature of trends and multimodal signals in social media.

Deep learning allows a prediction of viral videos using structured metadata such as views, likes, engagement ratios, and publish times in addition to unstructured text data such as titles, tags, and textual descriptions. This project uses the largest public YouTube trending dataset.

DATASET DESCRIPTION

Source: Kaggle , YouTube Trending Videos Global Dataset
Total Rows: 6,306,748 individual video trend records
In addition to the United States, it is also available in India, Japan, South Korea, Germany, Canada, the United Kingdom, Mexico, the Philippines, Brazil and Chile, representing multicultural content consumption and influencer behavior.

These fields capture the semantic and descriptive characteristics of each video:
• video_title , the headline text used to entice viewers.
• video_description , includes the full description, context, keywords, hashtags, and external links.
• video_tags , comma-separated tags assigned by the uploader, used for categorization or search optimization.
• These text components are important for audience discovery and influencer marketing performance, making them key areas for natural language modeling.

Structured numerical and categorical variables that describe how the video is performing and where it appears:
• video_view_count , number of views when the video became trending.
• video_like_count , total likes.
• video_comment_count , total comments.
• video_category_id , The category-specific mapping for YouTube's categories (e.g. Music, Entertainment, News).
• video_published_at , timestamp when the video was published.
• video_trending_country , the country in which the video appeared on the trending page.
• Metrics include platform engagement, viewer behavior, and algorithmic visibility.

Multiple new variables were created to improve model interpretability and performance:
• title_len - the length of the video title in characters.

- desc_len , length of description, in characters.
- tags_count , number of tags used.
- These features quantify creator optimization approaches and content richness.

These normalize engagement across videos with different audience sizes:
likes_per_view = likes / views
comments_per_view = comments/views
These are strong indicators of engagement and can even help to adjust for inflated raw view counts.

From the publication timestamp:
- publish_hour , hour of the day the video was uploaded.
- publish_dow: day of the week (0 for Monday, 6 for Sunday).
- Due to regional viewership patterns and YouTube's algorithm, upload time can considerably impact a video's reach on the platform.

Thus, this dataset is extremely large, and allows for deep learning models to be trained.
It covers multiple countries, accounting for variation in virality.
It also spans structured and unstructured data, making it suitable for multimodal deep learning.
It follows the rise of influencers on one of the world's largest video platforms.

METHODOLOGY:

• Data Cleaning & Preprocessing
A detailed preprocessing pipeline was devised to derive high-quality inputs to the deep learning models.

All rows with missing numeric fields (views, likes, comments) have been removed.
Missing entries in text fields (titles, tags, and descriptions) were replaced with empty strings.
Publish dates and trending dates were normalized to a common datetime format.

View counts, like counts, and comment counts were converted from object/string types to numerics.
Ratio features (likes_per_view and comments_per_view) were computed with a small epsilon term in order to avoid undefined behavior in the case of division by zero.
To prevent columns with larger ranges, such as views, from dominating the model, the numeric columns were standardized using StandardScaler.

To capture the full semantic and contextual meaning of each video:
video_title, video_tags, and video_description were represented as a single text field.
By concatenating textual input in this way, the models can learn from all the cues at once.

Keras Tokenizer was used to tokenize combined text.
Its vocabulary size has a limit around 50,000 tokens. These tokens represent the most frequent words and the most meaningful words.
All input sequences were padded or truncated to a common length of 200 tokens toward ensured consistency across models.

The columns with categorical features video_category_id video_trending_country received one-hot encoding.

Creates binary indicator features by category and country of trending appearance.

The engineered numeric and one-hot categorical variables were combined into a structured metadata matrix X_meta.
This matrix was also used in the MLP and Fusion models.

- **Train–Validation–Test Split**
  Therefore, to avoid data leakage in our evaluation, we use the following strict rules.

The dataset was split into:
- 80% Training Set: used for model training.
- 10% Validation Set - used for hyperparameter tuning and preventing overfitting.
- 10% Test Set - The 10% test set is held out until final evaluation.

The viral label was stratified and applied.
This preserved a real-world viral vs non-viral distribution of ~10% viral and ~90% non-viral.
It prevented the model from being biased towards the majority class.

The metadata matrices and text sequences were split in parallel for matching samples across modalities.
The respective MLP, LSTM, CNN, and Fusion models all utilized the same train/validation/test partitions.

To prevent data leakage, the metadata was scaled after the training/test split.
Tokenizer was only fit on the training text to simulate real-world prediction conditions.

MODEL ARCHITECTURES

We compare four types of deep learning models: one based only on the metadata, the other based only on the text and two models combining the two modalities using different fusion strategies. This allows to study the role of structured and unstructured features on prediction.

- **Multilayer Perceptron (MLP)**

In contrast, the MLP model is solely based on the structured metadata features, e.g. view counts, engagement ratios, category, time.

Numerical features standardized and metadata features one-hot encoded
* Dense Layer 1: 256 units, activation with ReLU
* Batch Normalization: Normalize training
* Dropout (0.4): Reduces overfitting in models
* Dense Layer 2: 128 neurons activate by ReLU
* Batch Normalization and Dropout
* Dense Layer 3: 64 units, ReLU activates
* Output Layer: 1 unit, Sigmoid (binary viral predication)

Due to the non-linear properties of the MLP, the model can learn associations of certain metadata with virality (e.g. post time, engagement ratio).

- **Long Short-Term Memory (LSTM) — Text Only**

The LSTM model processes the text from videos title tags and description It learns sequences and context.

* Embedding Layer: Maps words into 128-dimensional vectors
* LSTM Layer: 128 units. It captures long-range textual dependencies.
* Dropout: Prevents overfitting
* Dense Layer: ReLU Activates with
* Output Layer: Sigmoid activates

LSTMs are well suited to capture sequential patterns within language, such as the emotional tone of a video, the presence of clickbait, and the topic of a video.

- **Convolutional Neural Network (CNN) — Text Only**

The CNN model extracts local n-gram features from the text using short meaningful segments.

* Embedding Layer: Converts tokens to dense vectors.
* Conv1D Layer has 128 Filters. The Kernel Size is 5.
* Max Pooling: Highlights prominent features while reducing dimensionality
* Global Max Pooling: Captures strongest activating features
* Fully connected layer activates with ReLU
* Output Layer: Sigmoid activation

CNNs can be used to identify local patterns in text, such as clusters of keywords or influencers' language, that indicate trending discourse.

- **Multimodal Fusion Model — Metadata + Text (Best Performer)**

In addition to structured metadata, the Fusion model incorporates unstructured text to create a thorough representation of each video.

Branch 1 , Metadata Encoder

* MLP architecture (Section 5.1)
* Learns numerical and categorical relationships

Branch 2 , Text Encoder

* Embedding Layer
* Conv1D + Max Pool
* LSTM Layer
* Captures local patterns as well as long-term context

Fusion Layer

* Concatenation of metadata and text features
* Dense layers for combined representation
* Sigmoid output for final prediction

Virality is influenced by:

* Content semantics (text)
* Viewer engagement patterns (metadata)

The Fusion model outperforms the individual models across all metrics when they are combined which include accuracy, AUC, precision, recall, and F1-score (*Truong & Simmons, 2010; Kumar et al., 2017).*

RESULTS

Model	Accuracy	AUC	Percision	Recall	F1 Score
MLP	0.97	0.99	0.90	0.78	0.84
LSTM	0.98	0.97	0.77	0.71	0.74
CNN	0.94	0.97	0.79	0.66	0.72
Fusion	0.98	0.99	0.92	0.89	0.91

• *FUSION MODEL OUTPERFORMS ALL OTHER ARCHITECTURES*
The Fusion Model was the highest performing model across all metrics (accuracy, AUC, precision, recall, and F1 score).
This shows the clear value of multimodal learning, where the model utilizes both:
Structured metadata (such as engagement metrics, timestamps, categories)
Textual signals (titles, tags, and descriptions)
In integrating these sources of information, the Fusion Model is able to access both the content, as well as the audience's reception of that content, and predict virality more holistically.

• *METADATA ALONE IS SURPRISINGLY STRONG*
Even the MLP model with only metadata outperforms both the text-based models (LSTM and CNN).
This suggests that:
Engagement ratios (likes_per_view, comments_per_view)
Posting time
Category and country
View count dynamics
carry substantial predictive value.
In influencer marketing, behavioral analytics often provide more perception into the virality than text.

• *CNN AND LSTM PERFORM SIMILARLY*
The overall accuracy and AUC of these two text-based models are statistically comparable, with some trade-offs:
This makes the CNN more precise, as it avoids false positives by generating viral predictions only when certain.

When compared to the customary approach, the LSTM gives better recall (meaning more true viral videos are captured) but has more false positives.

These patterns are consistent with their architectures:

CNNs exploit local patterns of keywords in documents.

LSTM is particularly good at learning long-range dependencies.

- ***FUSION IMPROVES ALL METRICS***

In each case, the Fusion Model has the highest values.

Accuracy

AUC (near-perfect 0.9967)

Precision

Recall

F1 score

This indicates that:

Text cannot include every signal of virality.

A content's meaning cannot be determined by metadata alone.

Influencer performance can be best predicted using a model that combines semantic and behavioral features.

The findings show the importance of multimodal deep learning methods in influencer analysis on a large scale.

```
Predict on new video

    sample = "MrBeast goes crazy with a challenge"
    sample_seq = tokenizer.texts_to_sequences([sample])
    sample_pad = pad_sequences(sample_seq, maxlen=MAX_LEN)

    # metadata: just zeros placeholder
    dummy_meta = np.zeros((1, meta_input_dim))

    fusion_model.predict([dummy_meta, sample_pad])

1/1 ──────────────── 0s 124ms/step
array([[0.95905185]], dtype=float32)
```

To show the real-world applicability of the Fusion Model, a new sample title was selected.

"MrBeast goes crazy with a challenge."

The model scored each video based on the sigmoid, predicting 95.9 percent that the video would go viral. This is due to strong indicators like the creator's name and the use of impactful language, both of which are strong indicators that videos will go viral.

BUSINESS & MARKETING IMPLICATIONS

- *Predicting Viral Potential Before a Campaign Launch*

- For brands, being able to know when a piece of content is likely to go viral is essential to data-driven decision making.
- Marketers can identify patterns in metadata and textual information.
- Edit content drafts before publication
- Prioritize high-potential concepts
- Reduce campaign uncertainty
- Get better ROI from influencer partnerships
- This transforms influencer marketing from guesswork into a strategy.

• *Metadata Optimization Becomes a Strategic Requirement*
- This high model performance with only metadata also points to the importance of the creator-controlled variables:
- Posting time
- Video category
- Keyword tags
- Description richness
- Engagement ratios
- considerably predict viral outcomes.
- Creators and other organizations can use metadata features in the same way they would edit the video content itself, such as publishing their videos at the most effective time.

• *Identifying High-Potential Creators for Efficient Budget Allocation*
- Fusion-powered predictive models help brands make decisions:
- Choosing influencers for your campaigns
- Choosing the right budget to allocate
- Who consistently generates above average engagement
- Who is most likely to publish viral content
- Rather than obsessing over followers or other vanity metrics, marketers can invest dollars where the data suggests it will be effective.

• *Improving Recommendation and Ranking Algorithms*
- Multimodal models may be used by YouTube, TikTok and Instagram, for example.
- Improve ranking on the trending pages
- Improve personalized recommendations
- Surface high-quality content sooner
- Reducing algorithmic bias against larger creators
- By modeling both engagement behavior and semantic meaning of content, Fusion models can help platforms build a more accurate and fair discovery ecosystem.

• *Strategic Insights for Influencers and Agencies*
- Influencers can use these findings to:

- Create better titles, tags, and descriptions
- Post at optimal times
- Improve their content strategy based on engagement patterns
- Benchmark against predicted viral content
- Agencies may use it to:
- Data-driven predictions from pitch creators
- Improve campaign forecasting
- Utilize analytics for planned recommendations

CONCLUSION

This shows the potential of deep learning as a scalable framework for predicting the virality of influencer content in the real world. We show that by using a multimodal framework to consume both structured metadata and unstructured textual characteristics, we can predict the likelihood that a video will trend with very high accuracy. These results suggest that combining metadata and text is considerably better than either modality due to metadata describing how audiences interact with content and text describing the content, brand, and messaging strategy associated with it, creating a more thorough picture of brand engagement. These combined signals form a complete representation of the influencer's posts.

Apart from the high level accuracy in predictions achieved, this study has several practical implications for marketers, brands, and content creators. It can aid in marketing campaign planning, influencer budget allocation, determination of best social media post time, and content recommendation on content distribution platforms.

Future work can extend this work via similar embedding approaches such as thumbnail embeddings, perform sentiment analysis, or explore transformers-based reasoning such as Bidirectional Encoder Representations from Transformers (BERT), DistilBERT, and Text-To-Text Transfer Transformer (T5) models. Finally, with video frames, comments, or historical trend patterns fused, the system performs better and optimizes, allowing real-time viral forecasting.

REFERENCES

Dataset: *https://www.kaggle.com/datasets/canerkonuk/youtube-trending-videos-global*

Aaker, J. L., & Smith, A. (2010). *The dragonfly effect: Quick, effective, and powerful ways to use social media to drive social change*. Jossey-Bass.

Bengio, Y., Goodfellow, I., & Courville, A. (2016). *Deep learning*. MIT Press.

Chandrasekaran, R., de la Rosa, V., Kiehn, T., & Raghavan, H. (2023). Predicting online content virality using multimodal machine learning. *Journal of Marketing Analytics, 11*(2), 155–170.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *Proceedings of ICML*, 1597–1607.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Kumar, V., Choi, J., & Greene, M. (2017). Synergistic effects of influencer–brand fit and message appeal on consumer engagement. *Journal of Advertising Research, 57*(4), 370–385.

Liu, L., Lin, Z., Fang, X., & Li, S. (2021). Predicting social media popularity with multimodal deep learning. *Information Processing & Management, 58*(4), 102–109.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR).*

Truong, Y., & Simmons, G. (2010). Perceived intrusiveness in digital advertising: A consumer-centered approach. *Journal of Advertising, 39*(3), 37–51.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 5998–6008.

Zhang, X., & Zhao, M. (2020). Deep learning for online video virality prediction. *IEEE Transactions on Multimedia, 22*(10), 2607–2617.