YouTube Clickbait Classification using Video Transcript

Jay Ma

San Jose State University

Course Name: CS 180H

Professor Fabio Di Troia

May 18th, 2025

Abstract

The widespread use of captivating titles and eye-catching thumbnails on YouTube has led to the proliferation of clickbait—content that deliberately misleads viewers to increase viewership and generate revenue. This creates a disconnect between the video's presentation and its content, leading to user frustration, reduced trust in content creators, and the spread of misleading information. The core challenge is to accurately detect clickbait, as it can be subtle and context-dependent, often relying on linguistic cues and user engagement patterns. Traditional methods, such as manual flagging, are inefficient and subjective. Therefore, there is a need for automated, reliable systems that can identify clickbait based on multiple features, including textual content (titles, descriptions, transcripts) and metadata (likes, comments, etc.). This research aims to address this problem by experimenting with advanced machine learning models to accurately detect clickbait YouTube videos using transcripts.

Previous research uses title, description, and metadata information to perform binary classification of YouTube videos as either clickbait or not. However, the research does not take into account the transcript of the videos to train the models. This project aims to utilize video transcripts to make a more informed decision about the classification of YouTube videos as clickbait.

Contents

1 Introduction

Over the years, YouTube has evolved into one of the most influential and widely used platforms for video sharing, hosting a diverse range of content including tutorials, entertainment, commentary, and more. As of recent reports, there are approximately 37 million active channels on the platform, with nearly 500 hours of video content uploaded every single minute (Soundstripe, 2022). This unprecedented volume of content has led to intense competition among content creators who strive to capture and retain viewer attention in an increasingly saturated market.

To stand out, many creators resort to using captivating, often misleading, titles, and eye-catching thumbnails, and emotionally charged language to lure users into clicking on their videos. This is commonly known as clickbait which refers to content that intentionally misleads viewers by making exaggerated or deceptive claims in the title or thumbnail that the video itself does not align with. Clickbait tactics may drive short-term engagement, but they often lead to viewer dissatisfaction, frustration, and ultimately a loss of trust in the platform and its creators. Over time, such practices can harm the credibility of the entire platform and reduce user retention.

Given these consequences, the ability to automatically detect and filter clickbait videos has become an important challenge for content moderation and platform integrity. However, detecting clickbait is far from straightforward. The distinction between misleading and legitimate content is often subtle, relying on nuances in language, tone, and context. Traditional rule-based systems struggle to effectively capture this complexity, as clickbait does not always follow clear patterns.

Previous research has approached this problem by extracting features such as video titles, descriptions, thumbnails, likes, views, and comments—data that is available before a user clicks on the video (Gothankar et al., 2021). While such metadata can provide useful signals, these methods often overlook the actual content of the video, which plays a crucial role in determining whether the viewer's expectations were met. A title that seems exaggerated may be justified by the content, while a mundane title could still mislead in context.

This research aims to bridge that gap by focusing on the video transcript, which offers a more comprehensive and context-rich perspective of what the video truly delivers to the viewer. Unlike previous work, such as the master's thesis (Gothankar et al., 2021), which may

have incorporated transcripts alongside other features, this study uniquely focuses on the transcript as the sole input for predicting clickbait. This approach isolates the linguistic and contextual cues within the spoken content, providing insights into the potential of transcripts as standalone predictors.

To accomplish this, we explore a range of machine learning techniques, including logistic regression, random forest classifiers, and state-of-the-art transformer-based models such as DistilBERT. These models enable us to capture both surface-level and deep contextual representations of the transcript data. Through comparative experiments and evaluation, this study demonstrates the viability and effectiveness of using transcript-only models for the task of clickbait detection and highlights the advantages of content-focused analysis over purely metadata-based approaches.

2 Dataset

The dataset used for this research was found through the previous research (Gothankar et al., 2021), where a list of video IDs was obtained from a GitHub source (Vierti, 2022). Using the Google YouTube API, a python script was written to scrape the transcripts of clickbait and non-clickbait videos based on the IDs. A challenge was encountered during this process where many of the videos listed on the GitHub source were no longer available to extract the transcript or the transcript itself was not available. As a result, only a fraction of transcripts from the listed videos were retrieved. Specifically, about **18,000** clickbait and **19,000** non-clickbait video IDs were listed while only about 1,100 clickbait and 7,500 non-clickbait transcripts were retrieved. Additionally, an extra 70 transcripts were extracted based on 100 clickbait video IDs that was found in a dataset on kaggle (thelazyaz, 2021).

2.1 Preprocessing

In this research, the only feature we use to classify YouTube videos as clickbait or not is transcripts. Although transcripts can be used without any preprocessing, we still clean the trascripts used in research to train our models by removing punctuation. We also eliminate stop words using TF-IDF. Cleaning the transcripts in this manner helps improve model performance since punctuation and stop words such as "the" or "and" are simply noise to the machine learning models.

3 Experiments & Results

In these experiments, I tested multiple models with different embedding techniques after balancing the dataset. I also tested on unbalanced data set (the full dataset) which in turns all models performed badly. The experiments for the balanced dataset are listed and explained in detail below.

3.1 Word Embedding with TF-IDF

Since we use text data in this research, we first convert them into embeddings so that it can be used by the machine learning algorithms for training. The first word embedding technique we used to vectorize the transcripts is the TF-IDF (Term Frequency-Inverse Document Frequency) in scikit-learn. TF-IDF is a common tool used to vectorize the text data, it works by evaluating how important a word in a corpus depending on how frequently it occurs. Rare and document-frequent words are given more importance by TF-IDF.

We evaluate the performance of Random Forest, Multi-Layer Perceptron (MLP), and Logistic Regression (LR) in binary classification of transcripts (vectorized by TF-IDF) as clickbait or non-clickbait. For reference in the figures and tables below, **non-clickbait is encoded as 0 and clickbait is encoded as 1**. The results of each model as confusion matrices and classification reports are shown below:

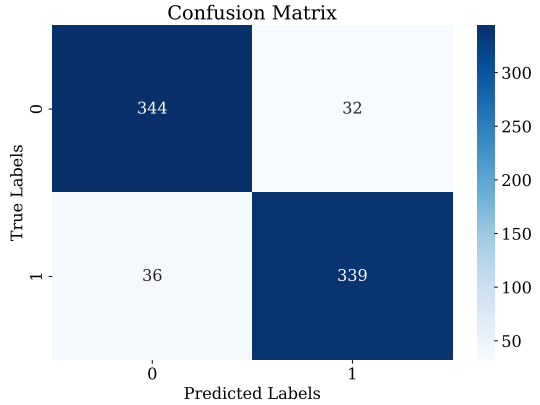

Figure 1

TF-IDF Random Forest Confusion Matrix

Table 1

TF-IDF Random Forest Classification Report

	precision	recall	f1-score	support
0	0.91	0.91	0.91	376.00
1	0.91	0.90	0.91	375.00
accuracy	-	-	0.91	751.00
macro avg	0.91	0.91	0.91	751.00
weighted avg	0.91	0.91	0.91	751.00

The confusion matrix and classification report above show that Random Forest trained with 100 estimators on TF-IDF encoded transcripts is accurately classify them with a weighted average **precision of 0.91** and **recall of 0.91**. Random Forest offers the best

performance observed in the experiments.

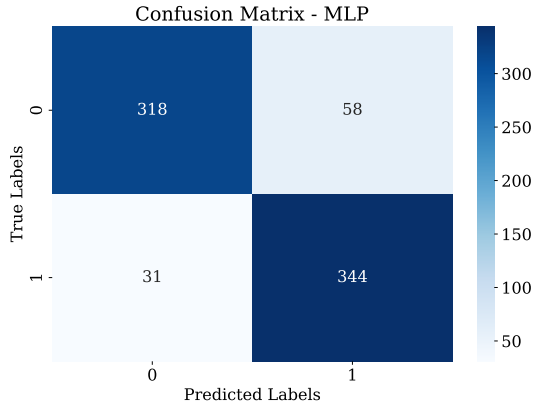

Figure 2

TF-IDF MLP Confusion Matrix

Table 2

TF-IDF MLP Classification Report

	precision	recall	f1-score	support
0	0.91	0.85	0.88	376.00
1	0.86	0.92	0.89	375.00
accuracy	-	-	0.88	751.00
macro avg	0.88	0.88	0.88	751.00
weighted avg	0.88	0.88	0.88	751.00

The MLP from scikit-learn is also able to accurately classify TF-IDF encoded transcripts with a weighted average **precision of 0.88** and a **recall of 0.88**; however, it is still outperformed by LR (Table 3) and Random Forest.

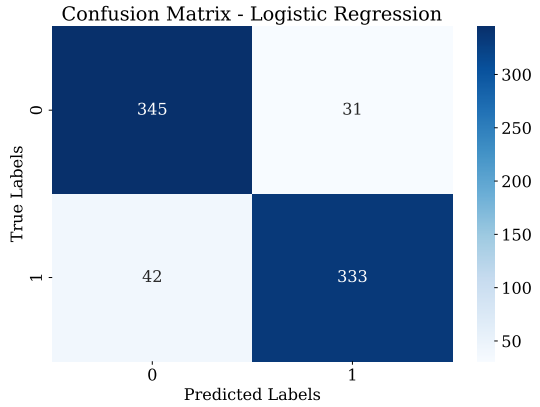

Figure 3

TF-IDF LR Confusion Matrix

Table 3

TF-IDF LR Classification Report

	precision	recall	f1-score	support
0	0.89	0.92	0.90	376.00
1	0.91	0.89	0.90	375.00
accuracy	-	-	0.90	751.00
macro avg	0.90	0.90	0.90	751.00
weighted avg	0.90	0.90	0.90	751.00

Although a simple and interpretable model, LR shows great performance in classifying transcripts encoded with TF-IDF and achieves a weighted average **precision of 0.90** and **recall of 0.90** (shown above in Table 3 and Figure 3) which is almost on par with the more rigorous Random Forest Classifier and outperforms MLP by a small margin of 0.02.

3.2 Word Embedding with Word2Vec

Word2Vec is another popular embedding technique first introduced in 2013. Word2Vec is a neural network based approach that learns vector representations (embeddings) of words based on their context in a corpus. In essence, words appearing in similar contexts have similar vectors when embedded using Word2Vec.

We evaluate our models again and the results of each models as a confusion matrix and classification report are shown below:

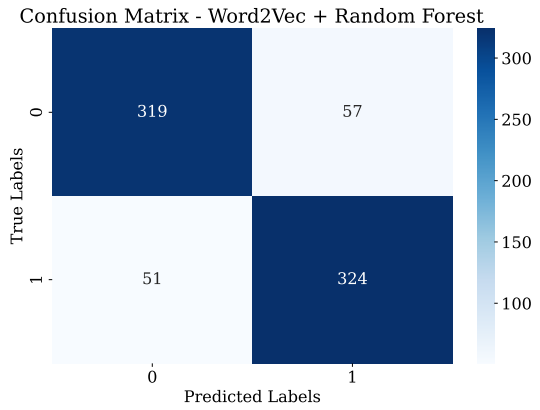

Figure 4

Word2Vec Random Forest Confusion Matrix

Table 4

Word2Vec Random Forest Classification Report

	precision	recall	f1-score	support
0	0.86	0.85	0.86	376.00
1	0.85	0.86	0.86	375.00
accuracy	-	-	0.86	751.00
macro avg	0.86	0.86	0.86	751.00
weighted avg	0.86	0.86	0.86	751.00

By comparing the classification report in Table 1 for Random Forest with TF-IDF encoded transcripts to the report in Table 4 for Word2Vec encoded transcripts, we can observe a decrease in performance. The model performs better with TF-IDF than with Word2Vec in this case. Random Forest trained on TF-IDF encoded transcripts outperforms its Word2Vec counterpart by **0.05**.

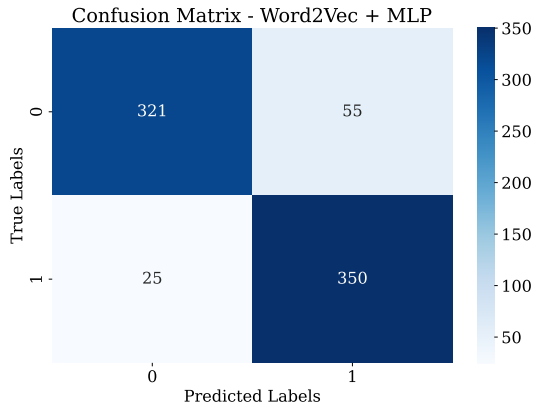

Figure 5

Word2Vec MLP Confusion Matrix

Table 5

Word2Vec MLP Classification Report

	precision	recall	f1-score	support
0	0.93	0.85	0.89	376.00
1	0.86	0.93	0.90	375.00
accuracy	-	-	0.89	751.00
macro avg	0.90	0.89	0.89	751.00
weighted avg	0.90	0.89	0.89	751.00

Unlike Random Forest and Logistic Regression, MLP is the only model that increased in performance with Word2Vec encoded transcripts. MLP increased performance by about **0.02** in both weighted average precision and recall. MLP now outperforms both Random Forest and Logistic Regression by **0.04** and **0.03** in f1-score, respectively.

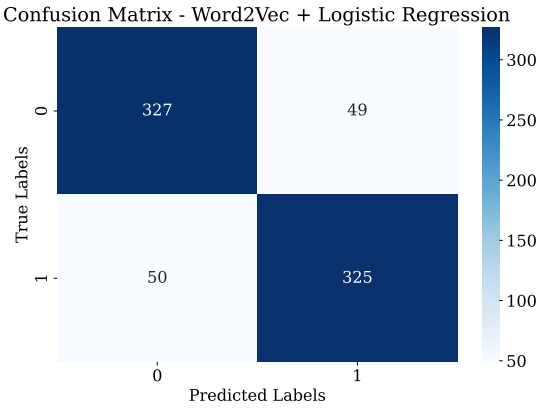

Figure 6

Word2Vec LR Confusion Matrix

Table 6

Word2Vec LR Classification Report

	precision	recall	f1-score	support
0	0.87	0.87	0.87	376.00
1	0.87	0.87	0.87	375.00
accuracy	-	-	0.87	751.00
macro avg	0.87	0.87	0.87	751.00
weighted avg	0.87	0.87	0.87	751.00

Similar to Random Forest, we can notice in the above Table 6 that LR has decreased in weighted average precision and recall by **0.03** compared to its TF-IDF counterpart.

3.3 Embedding using DistilBERT

DistilBERT is a transformer-based embedding technique that builds on BERT (Bidirectional Encoder Representations from Transformers), which was introduced by Google in 2018. DistilBERT is a lighter and faster version of BERT while being more efficient to run.

It generates contextualized word embeddings by considering the entire sentence, allowing each word's meaning to be influenced by the words around it. This makes DistilBERT particularly effective for tasks involving nuanced language understanding. In this research, we use DistilBERT to convert transcripts into dense vector representations that capture semantic relationships between words and phrases.

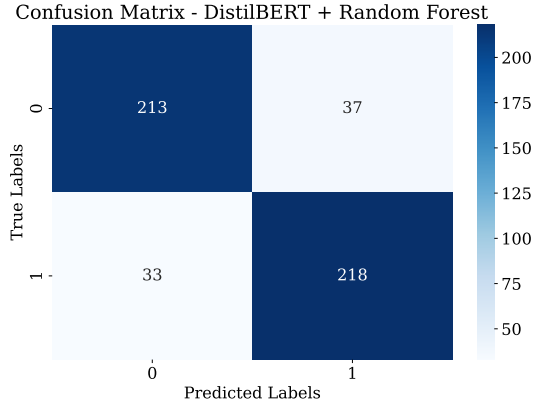

Figure 7

DistilBERT Random Forest Confusion Matrix

Table 7

DistilBERT Random Forest Classification Report

	precision	recall	f1-score	support
0	0.87	0.85	0.86	250.00
1	0.85	0.87	0.86	251.00
accuracy	-	-	0.86	501.00
macro avg	0.86	0.86	0.86	501.00
weighted avg	0.86	0.86	0.86	501.00

The Random Forest model with DistilBERT embedding has similar performance to Word2Vec embedding (Table 4). Both embeddings have a weighted average **precision of 0.86** and **recall of 0.86**. However, it still falls short of the TF-IDF model (Table 1), which outperforms both Word2Vec and DistilBERT with a higher precision and recall of **0.91**.

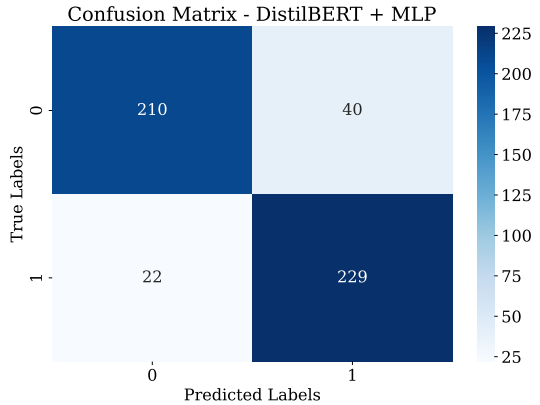

Figure 8

DistilBERT MLP Confusion Matrix

Table 8

DistilBERT MLP Classification Report

	precision	recall	f1-score	support
0	0.91	0.84	0.87	250.00
1	0.85	0.91	0.88	251.00
accuracy	-	-	0.88	501.00
macro avg	0.88	0.88	0.88	501.00
weighted avg	0.88	0.88	0.88	501.00

The MLP model with DistilBERT embedding achieves a weighted average precision and recall of **0.88**, matching the performance of MLP with TF-IDF embedding (Table 2). However, unlike the MLP model with Word2Vec embedding (Table 5) achieving the highest performance, DistilBERT does not lead to further improvement.

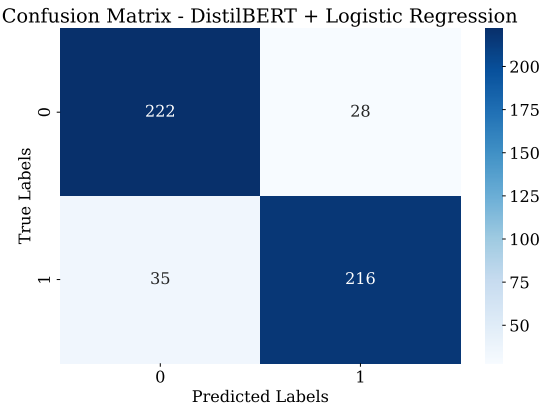

Figure 9

DistilBERT LR Confusion Matrix

Table 9

DistilBERT LR Classification Report

	precision	recall	f1-score	support
0	0.86	0.89	0.88	250.00
1	0.89	0.86	0.87	251.00
accuracy	-	-	0.87	501.00
macro avg	0.87	0.87	0.87	501.00
weighted avg	0.87	0.87	0.87	501.00

The Logistic Regression model with DistilBERT embeddings achieves a weighted average **precision and recall of 0.87**, which is consistent with the performance of the same model using Word2Vec embeddings (Table 6). However, the DistilBERT-based model has an **underperformance of 0.03** for the weighted average of precision and recall compared to TF-IDF (Table 3)

4 Conclusion

This study demonstrates the effectiveness of using video transcripts as a standalone feature for classifying YouTube videos as clickbait or non-clickbait. By applying various machine learning models with embeddings such as TF-IDF, Word2Vec and DistilBERT, we found that transcript-based features can yield high classification accuracy, with Random Forest and Logistic Regression performing best with TF-IDF, while MLP showed slightly improved results with Word2Vec. These findings highlight the potential of leveraging spoken content for clickbait detection, offering a more content-aware and scalable alternative to traditional metadata-based approaches. However, we must also keep in mind the limitations of this study. The dataset used for this research was small due to the challenges and intense manual labor involved in collecting transcripts. Therefore, the dataset may not represent the actual population of clickbait videos in video-sharing platforms. Future work should include a

significantly larger dataset with transcripts preferably from various video-sharing platforms to ensure diversity in transcript content. In addition, a remedy for the roadblocks faced while collecting transcripts and expanding the dataset can be training a Generative Adversarial Network (GAN) model. This model can be used to generate realistic clickbait transcripts that augment existing small datasets which can help train models on more diverse data.

References

Gothankar, R., Di Troia, F., & Stamp, M. (2021). Clickbait detection in youtube videos. *arXiv preprint arXiv:2107.12791*. https://arxiv.org/abs/2107.12791

Soundstripe. (2022). How to beat the YouTube algorithm and get your videos to rank in 2022. https://www.soundstripe.com/blogs/how-to-beat-the-youtube-algorithm-and-get-your-videos-to-rank-in-2022

thelazyaz. (2021). Youtube clickbait classification. https://www.kaggle.com/datasets/thelazyaz/youtube-clickbait-classification

Vierti, A. (2022). Youtube-clickbait-detector. https://github.com/alessiovierti/youtube-clickbait-detector