Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/4/20230435

Commercial video recognition system for short video (TikTok) based on machine learning

Mingyuan Fang

Department of Computer Science, University of Toronto, 40 St. George Street, Toronto, ON, M5T 2E4

mingyuan.fang@mail.utoronto.ca

Abstract. Short video has the features of short duration and high information carrying capacity, which is more in accordance with modern netizens' mobile phone using patterns. With the continual increase of the user scale of smart mobile terminals, many mobile phone users may make full use of the fragmented time to shoot and view short movies. Numerous Internet behemoths are fighting to invest in creating short video platforms since the amount of video user traffic generates enormous commercial prospects. For speeding up the audit team's effectiveness, video classification technology needs to be constantly developed and updated. The article proposed a commercial video detection model with a wide range of data analysis and processing. More specifically, Principal Component Analysis (PCA), feature selection by random forest and discretization using decision trees would be involved in order to transform the original data into features that better express the nature of the problem. The application of these features to Random Forest Model can improve the model prediction accuracy of data. Experimental results demonstrate that the recognition system fulfills outstanding performance. The model achieves 0.90 precision and 0.96 AUC score (area under ROC curve) of excellent evaluation in the corresponding test set.

Keywords: Classification, TikTok, Principal Component Analysis (PCA), Random Forest.

1. Introduction

Short video, particularly for the TikTok platform, is a type of Internet content dissemination method that typically refers to the Internet new media distribution of video material lasting little more than 5 minutes. Short, quick, and large-flow communication material is steadily winning over important platforms, admirers, and funding because to the growing use of mobile devices and the network's accelerated growth. In order to fight problematic videos, short video platforms began to prioritize content filtering. On the basis of extending the audit team and strengthening the review mechanism, relevant short video platforms encourage the use of machine learning, recognition, and other technologies in the content review process to increase the accuracy and coverage of the review. The article is concerned with one sort of video that regularly causes non-standard difficulties — commercial video; the goal of the study is to aid platforms in separating commercial video from mass distribution video in terms of machine learning, which speeds up the audit team's efficiency.

In a number of fields, machine learning has made enormous strides [1-5]. However, there hasn't been much focus on the intersection of machine learning and social online media, especially for self-

DOI: 10.54254/2755-2721/4/20230435

media. The Internet and the ongoing evolution of social network services, according to [6], have flattened the earth. In order to determine the veracity of news, the author suggests a bottom-up approach with relative, mutual, and dynamic credibility assessment using a dynamic relational network (or mutual evaluation model), where each node may evaluate its own credibility and be evaluated in turn by other nodes based on the coherence of its content. Many people spread false information on social media. On the development of web media users' rights protection, one study analyze the important factors that affect We-media video producers' desire to secure their rights in terms of machine learning algorithms in order to comprehend how self-media video consumers see the need for copyright protection for original videos and the future direction of the self-media video ecosystem [7]. Regarding the advancement of self-media data learning, the author noted that the issue of word vector ambiguity lowers the precision of intent identification [8]. Additionally, there is a divergence between local and global features in the meanwhile, which prevents text feature extraction from accurately reflecting semantic information. All these problems are obstacles to intention detection. Therefore, an attention-based convolutional neural network (A-CNN) for marketing intention is proposed for learning from self-media data. The high-dimension word vectors in the A-CNN are represented by a quick feature extraction method based on skip-gram-based learning called FSLText, which is also proposed by [3]. Attention networks are then created by cascading the traditional CNN with the self-attention model, resulting in the creation of the novel network structure known as the A-CNN. From the perspectives of the categorization of social media users, one study utilized with Twitter real dataset; Individuals, the media, the government, and organizations are four different categories of social media users [9]. Four types of users may be precisely identified by the classification model based on support vector machines (SVM) and stochastic gradient descent (RGDESCENT). It serves as a point of reference for the upcoming study on the classification of social media users. However, the self-media in the aspects of short video classification is rarely studied.

People are increasingly using short movies to get information. Relevant platforms like TikTok must advance machine learning research in this area in order to increase productivity and quality while preserving a positive atmosphere for online media. Therefore, for we media to further improve the short video environment, fresh short video categorization algorithms must be proposed. This study combines the decision tree and random forest models for the model selection. To increase the effectiveness of the training model, a variety of procedures are utilized in data preparation, including data cleaning, feature filtering, PCA feature creation, feature binning via decision trees, and feature significance ranking. Details may be found in the Method section.

2. Method

In-depth discussion of the experiment's methodology and its interpretation will be provided in this section. A description of the dataset, an explanation of the data pretreatment steps, and an analysis of the machine learning algorithm employed are all included.

2.1. Data description

Advertising recognition is an issue in semantic video classification. The audio-visual content that makes up an advertisement on a specific platform is distinguished by an audio-visual presentation. As a result, unique audio-visual components must be taken out of the video clip. TikTok video poses challenging machine learning challenges since it lacks a specific video style, is very unpredictable, and is dynamic.

The AD dataset comprises common audiovisual elements of video footage taken from TikTok video, with five samples and a total of 150 hours of video, captured at 270 frames per second and 720 x 576 resolution. To determine if a video contains commercial advertising, the video data is processed into the characteristics of the video time, sound spectrum, video spectrum, text distribution, and picture change. One label and 230 features are present in the final data. The source of the dataset is from TIANCHI [10].

DOI: 10.54254/2755-2721/4/20230435

The commercial_video_data.csv file contains information on 129,685 movies with labels denoting labels. There are 230 features in the dataset, including variations in picture quality, text distribution, sound spectrum, and video spectrum. The unit for producing the sample is the video shot. RGB color histograms are used in video to separate subsequent video frames into individual video shots. The dataset recorded the fundamental frequency, the short-term energy and zero crossing rate, the spectral centroid, the spectra flux, the spectrum roll-offs frequency, and the MFCC audio word package for each video camera. It also recorded five visual features (i.e., video lens length, each video camera screen text distribution, distribution, the frame difference distribution, edge rate).
Overall, the type of task is binary classification. Label 1 denotes commercial advertising whereas label -1 represents non-advertising.

2.2. Data preprocessing

Data preprocessing, an important step in the data analysis process, may be described as the modifying or deleting of data before to use in order to ensure or improve performance. This section could be categorized as Data Observation and Cleansing, Feature Engineering, and Machine Learning model.

2.2.1. Data observation and cleansing. Importing the dataset and necessary modules (i.e., pandas) is the first step. Next, rename a few crucial features based on Pandas. The dataset looks like this following these processes: Implementing the seaborn function draws a histogram of the sample distribution. The following diagram shows that Label = -1 represents non-advertisements, Label = 1 represents advertisements and the number of advertisements is relatively large in this data set.

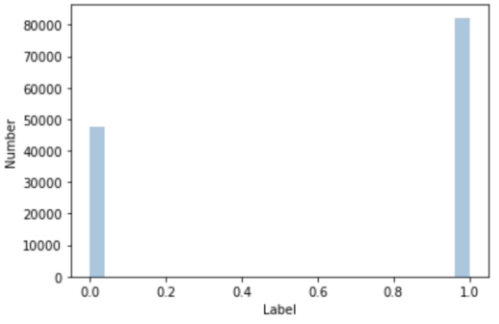

Figure 1. Data distribution.

Rearranging Label -1 to 0 for convenience. Figure 1 shows the data distribution.

To examine the feature distribution in the data, the distribution, and statistics of the significant feature Length are then drawn shown in Figure 2 and Figure 3.

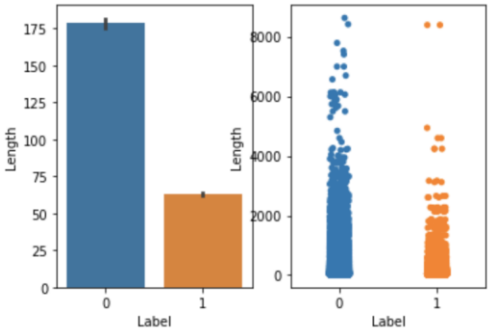

Figure 2. Length distribution in labels.

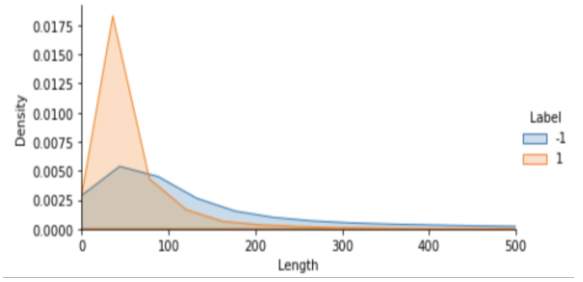

Figure 3. Length density distribution.

The initial dataset may include NaN, thus the second step is to clean it up. To determine which column has NaN value; we use the corresponding functions provided by pandas to search, to fill in the missing data, and to eliminate duplicate samples. After, The shape of data was altered from the initial (129685, 231) to (111615, 231).

DOI: 10.54254/2755-2721/4/20230435

2.2.2. Feature engineering. For feature engineering, many techniques can be combined. First, significant characteristics are selected using the Random Forest approach. The second is the application of discretization to improve data stability and lower the likelihood of overfitting. Thirdly, based on the PCA algorithm, reducing the dimension and complexity of the dataset.

Feature selection by random forest. Random forest is employed for feature selection. The random forest model was fitted to the training set for obtaining feature_importances_ (the attribute of Random Forest Object). Plotting the bar chart by order of corresponding feature_ importances_; Calculate the total of the important values to plot the ladder shown in Figure 4. Further details of the random forest model will be discussed in section 2.3.

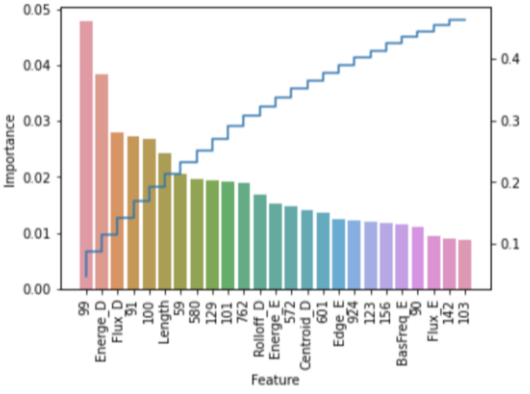

Figure 4. Feature ranking diagram.

Feature binning by decision tree. A continuous or numerical variable is encoded into a categorical variable via binning or discretization. Non-linear models can occasionally struggle with numerical or continuous features. Therefore, binning continuous variables brings non-linearity into the data and tends to enhance model performance. The process of discretizing data involves using a decision tree to find the best dividing points that would create the bins or continuous intervals: Step 1: The variable we wish to discretize is used to train a decision tree with a limited depth to forecast the target. Step 2: Next, the probabilities the tree returned are substituted for the values of the original variables. All the data contained in a single bin have the same probability, therefore replacing by probability is equal to grouping the observations according to the cut-off value selected by the decision tree.

Dimensionality reduction by PCA. Principal Component Analysis (PCA), one of the most widely used unsupervised machine learning techniques, is used in a variety of applications, including data exploration, dimensionality reduction, information compression, data de-noising, and numerous others. Because PCA is computed by identifying the components that account for the most variance, which captures the signal in the data and leaves out the noise, the approach is used to reduce the number of dimensions in the training dataset and de-noise the data when it is used for preprocessing. The result is a dataset that highlights the most crucial aspects of the data. Finding the optimal dimensionality is necessary before you can truly lower it. We find suitable n_components (the number of features remaining) through visualization of explained_variance_ratio (the attribute of PCA object) based on the current number of features. After getting the n_component, we use both training set and test set to fit the PCA model in terms of fit_transform function respectively, and thus obtaining the corresponding sets after dimensional reduction. Figure 5 shows the explained variance ratio.

DOI: 10.54254/2755-2721/4/20230435

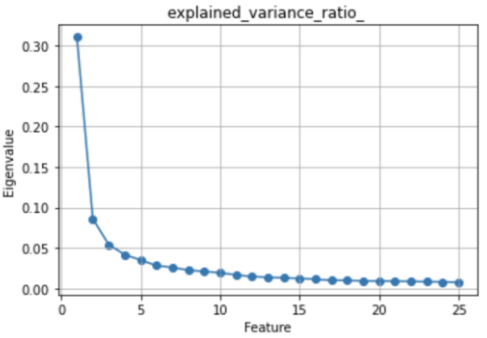

Figure 5. Explained variance ratio.

2.3. Modeling and parameter tuning

In the choice of modeling, the Random Forest model was selected. A random forest, as its name indicates, is a collection of various independent decision trees that function as an ensemble. The class that obtains the most votes become the model's prediction. Each tree in the random forest puts out a class forecast. The Random Forest Algorithm's capacity to handle data sets with both continuous variables, as in regression, and categorical variables, as in classification, is one of its significant features. It generates superior outcomes for classification problems. For the parameters setting of the RandomForestClassifier model: max_depth = 12, max_features = 16, n_estimators = 2000, n_jobs = -1, random_state = 0 tend to be set.

3. Result and discussion

To analyze the result, 2 types of evaluation methods were implemented, which is Prediction Accuracy and AUC.

Table 1. Performance of the model.

Type	Train Set	Test Set
Accuracy Score	0.924765	0.899226
AUC Score	0.978776	0.958072

3.1. Prediction accuracy

The Table 1 shows that the model achieved a high degree of prediction accuracy, which proves the effectiveness of the model in this study. In the training set and test set, the prediction accuracy was as high as 0.92 and 0.89, respectively.

3.2. Area under curve

AUC, a model assessment indicator, evaluates the binary classification model in order to more accurately assess the outcomes and model. The AUC score in the training set, as shown in the table, reaches a high of 0.978, and it also performs well in the test set, reaching approximately 0.958. The ROC curve for the matching test set is also shown in Figure 6. The area under the ROC is around 0.96, as can be shown.

DOI: 10.54254/2755-2721/4/20230435

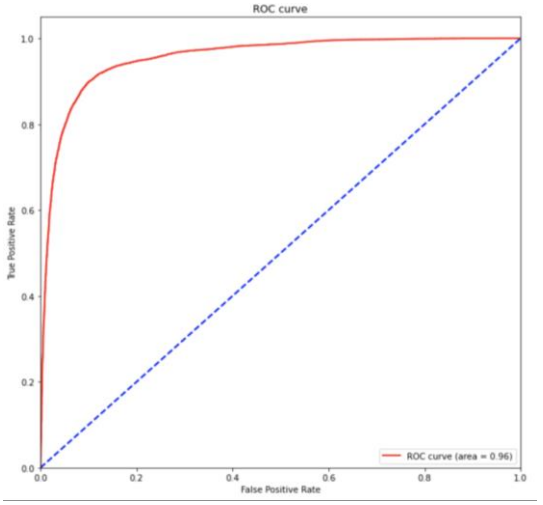

Figure 6. ROC curve.

4. Conclusion

For this work, the study's objective is to help platforms differentiate commercial videos from a large number of submitted videos using machine learning, which boosts the audit team's productivity. The article applied a variety of implementations e.g., feature selection method, discretization, PCA algorithm and Random Forest model in the steps of feature engineering and modeling to improve the ability of short videos classification. The comprehensive assessment methods were operated to evaluate the performance of the model. The result shows that the model, commercial video recognition system, has excellent outcomes on accuracy prediction and AUC. In the future, the system will be adapted from binary classification to be a multiclass classification model to serve the short video platform.

References

[1] Girshick R 2015 Fast r-cnn In Proceedings of the IEEE international conference on computer vision p1440-1448
[2] Qiu Y et al. 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72 103323
[3] He K et al. 2017 Mask r-cnn Proceedings of the IEEE international conference on computer vision
[4] Rigatti S J 2017 Random forest Journal of Insurance Medicine 47 31-39
[5] Biau G and Erwan S 2016 A random forest guided tour Test 25.2 197-227.
[6] Sshida Y and Sanae K 2018 Fake news and its credibility evaluation by dynamic relational networks: A bottom up approach Procedia Computer Science 126 2228-2237
[7] Wu F et al. 2020 Research on Self-media Original Video Protection Based on Machine Learning——Based on the original author's perspective Academic Journal of Business & Management 2.4
[8] Hou Z et al. 2021 Attention-based learning of self-media data for marketing intention detection Engineering Applications of Artificial Intelligence 98 104118
[9] Li G et al. 2019 Research on Social-media User Classification Based on Machine Learning Modern Library and Information Technology 003.008(2019) 1-9
[10] Tianchi Commercial Video Data 2020 https://tianchi.aliyun.com/dataset/dataDetail?dataId=53460