# Video Classification into Academic and Entertainment using Subtitles

**Raghav Agarwal[a], Manikanta P L V N P [b], Saad Yunus Sait[c*]**

[a]SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kanchipuram, Chennai, Tamil Nadu, India. E-mail: RAGHAV1999AGG@GMAIL.COM
[b]SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kanchipuram, Chennai, Tamil Nadu, India. E-mail: MANIKANTAPAVAN10@GMAIL.COM
[c*]SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kanchipuram, Chennai, Tamil Nadu, India. E-mail: SAADY@SRMIST.EDU.IN

**Article History**: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

**Abstract:** Automated classification of text into predefined categories has always been considered as a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing. This kind of web information is popularly known as the digital/electronic information. It is in the form of documents, conference material, publications, journals, editorials, web pages, videos, e mail etc. People largely access information from these online sources rather than being limited to archaic paper sources like books, magazines, newspapers etc. But the main problem is that this enormous information lacks organization which makes it difficult to manage. Text classification is recognized as one of the key techniques used for organizing such kind of digital data. This paper describes our work on building a Classification system for video subtitles. As an example, we report on our evaluation results for two TV genres -Academic and Entertainment. Through this implementation, the user can classify the videos into academic and entertainment in any software or system, which will help the user to differentiate the videos properly into useful and distracting content. This enables applications like social networking and instant messaging apps to filter distracting content, thereby enhancing productivity of users.

**Keywords:** Text-Classification, Bag-of-Words(BOW), Random-Forrest-Classifier, Data- Acquisition, Scikitlearn.

## 1. Introduction

Now-a-days the amount of information available on the web is tremendous and increasing at an exponential rate. Automatic text classification has always been an important application and research topic since the inception of digital documents to manage the enormous amount of data available on the web (Ikonomakis et al., 2005). It is based on machine learning techniques that automatically build a classifier by learning the characteristics of the categories from a set of pre-classified documents (Sebastiani, 2002). It plays an important role in information extraction and summarization, text retrieval, and question- answering. Typically, most of the data for classification is of heterogeneous nature collected from the web, through newsgroups, bulletin boards, and broadcast or printed news scientific articles, news reports, movie reviews, and advertisements. They are multi-source, and consequently have different formats, different preferred vocabularies and often significantly different writing styles even for documents within one genre. Therefore, automatic text classification is highly essential

Text classification is the task of classifying a document under a predefined category. More formally, if[Di] is a document of the entire set of documents[D] and [C1, C2, C2,..., Cn] is the set of all the categories, then text classification assigns one category [Cj] to a document [Di] (Ikonomakis et al., 2005). The documents depending upon their characteristics can be labeled for one class or for more than one class. If C= [0, 1], it is called a “Binary classification problem”. If C= [0, 1, 2, 3, 4... n], it is called a ‘multi classification problem” (WangandChiang, 2011). Using this a video can be classified into different categories using its subtitles as a text document and classifying the text document to a label.

This can be used in several ways such as having a huge collection of videos which consists of WhatsApp videos, Instagram videos, reels, movies, series, anime, lectures, course content videos etc and this is the best way to use them and classify them. This is done by making a YouTube playlist of academic and Entertainment videos and extracting information and creating the data set through a sysdrive python script. Then, BOW (Bagging of words) featurization can be implemented. Using these unique features, machine -learning models are trained using scikitlearn lib. Through this implementation, the user can classify the videos into Academic and

Entertainment in any software or system, which will enable the applications to handle both categories differently. More specifically, this can be used to filter distracting content on a social networking site.

GOOGLE2SRT tool directly converts the YouTube video link to the subtitle in a SRT format. Through this train and test data set is made after preprocessing and cleaning the SRT files and the examples are used to train the following 10 machine learning models: Logistic -Regression, K-NN, Decision Trees, Rigid Classifier, Gradient Bossting-Classifier, Random-Forrest, Bernoulli Naive-Bayes, Multinimial - Naive-Bayes, Complimental-Naive-Bayes and Gaussian-Naive-Bayes. Doing the cross validation testing and learning about the working of the models, Random-Forrest-Classifier was the final selection with 0. 856 F1 Score and due to its property of converting (Low-Bias - High-Variance) to (Low-Bias - Low-Variance) and thereby providing regularization.

## 2. Literature Review

The main purpose of the project is to filter out the distracting connect from a given set of videos by dividing them into academic and Entertainment. Text-Classification is a very researched field and many great publications and research papers are available for such problem statement. The papers mentioned in the Reference Section are the accumulation of testing papers and applied research papers where some papers talk about specific linguistic behavior of text while some check different models and their comparison for the task of classification.

Conversion of text into D--dimensional vector is the main property through which text can be categorized into different labels and hence this property of vectors of text tells how similar are some cluster of text. The main property these vector text have is similarity is inversely proportional to the distance, i.e. if the distance between the vectors is too large, then text tend to be less similar and vice versa.

For conversion many types of methods are used i.e.: Bag-of-words (BOW), TF-IDF, WORD-2-VEC, Bert-Embeddings. With respect to videos, there is always an alternative solution classifying frames of video into academic and Entertainment. This can be done using pre-trained VGG-NET model (Karen Simonyan and Andrew Zisserman - 2014) and apply transfer learning and train a Neural-Network on top of it, which can easily give great results.

The best text classification technique is the Bert-Transformer (Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova - 2018) where, the text corpus is converted to Bert Embedding and using these Bert embedding as a dataset to do the classification of text. This is also called Transfer learning of NLP tasks where we take the pre-trained weights of Bert-Transformer model and convert the text and using these embedding make a simple forward- Neural Network and classify the text.

In this paper, only simple models of classification have been considered, with an objective of reducing the footprint of the model on a device like a smartphone. So, BERT embeddings have not been considered.

Further, only classification of videos using subtitles rather than video frames has been considered.

For this implementation using text classification with simple ML-Algorithms was the goal, looking at videos related to academic and Entertainment, the first observation was that subtitles differs significantly; the simple and effective approach came out to be using BOW for featurizing and use of these unique features to classify the videos into academic and Entertainment videos using the subtitles of the video.

## 3. Experimental - Setup

### 3.1. Data Acquisition

The main source for getting the videos was YOUTUBE due to the wide variety of content due to which, the model will not tend to be biased in terms of national or foreign videos.

For extracting subtitles of the video from YOUTUBE, one option was to write a sysdrive script and point to the data needed and collect the subtitles, however, due to large variation in the location of the transcripts in YouTube HTML page, this was not performed; rather, the GOOGLE2SRT tool as used, in which the user has to provide a text document of links to the videos and for each link a SRT file will be saved in the destination folder. This helped in making more than 30K files in total. Following this approach, two text documents were generated, one each for academics and entertainment.

### 3.2. Data Processing and Analysis

An SRT file (SubRip Subtitle file) is a plain-text file that contains critical information regarding subtitles, including the start and end time codes of text to ensure subtitles match audio, and the sequential number of subtitles. But in our implementation, we were not taking in account the information passed in the sequence but rather were just concerned with the words in the text.

In the pre-processing and cleansing stage, two data frames -academic and Entertainment with columns “sub-titles” and “label” were generated, with use of folders created using GOOGLE2SRT tool, after cleaning the text into valid format and only considering English text for training the model.

After making the two data frames, the two data frames were concatenated and then simple shuffling using predefined shuffle function in pandas library on the rows is given to the data and the main DATAFRAME was ready.

The main motive of the Data acquisition was to have no imbalance in the main Data Frame and that was achieved by monitoring the number of files made by the GOOGLE2SRT tool in both the folders as shown in fig -1; it may be noted from Fig. 1 that the examples of both classes are balanced. It was also noted that some subtitles consisted of more than 50000 words as shown in Fig -2.

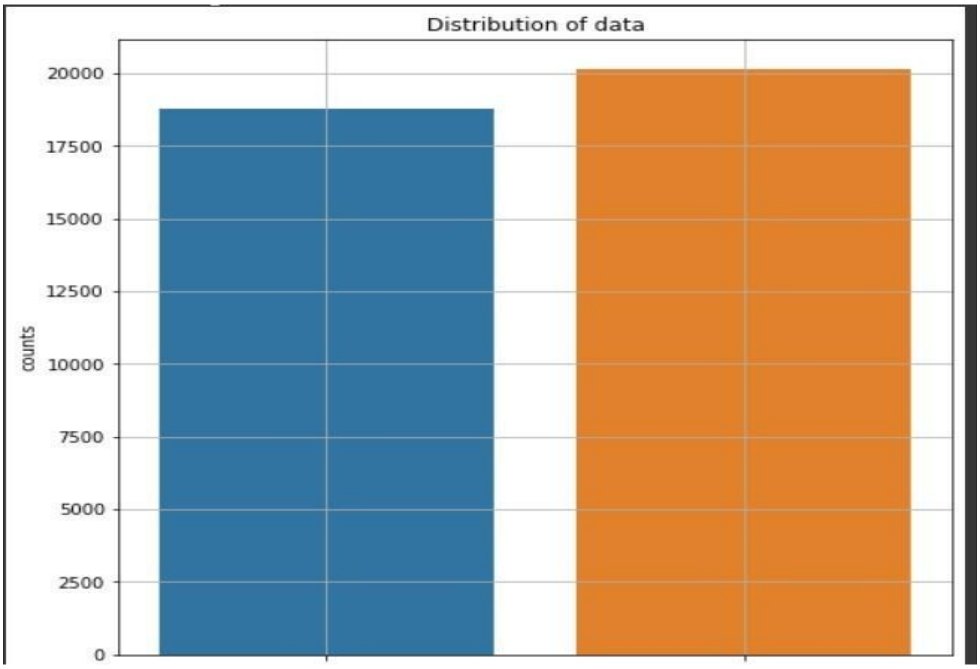


**Figure 1.** Distribution of the data

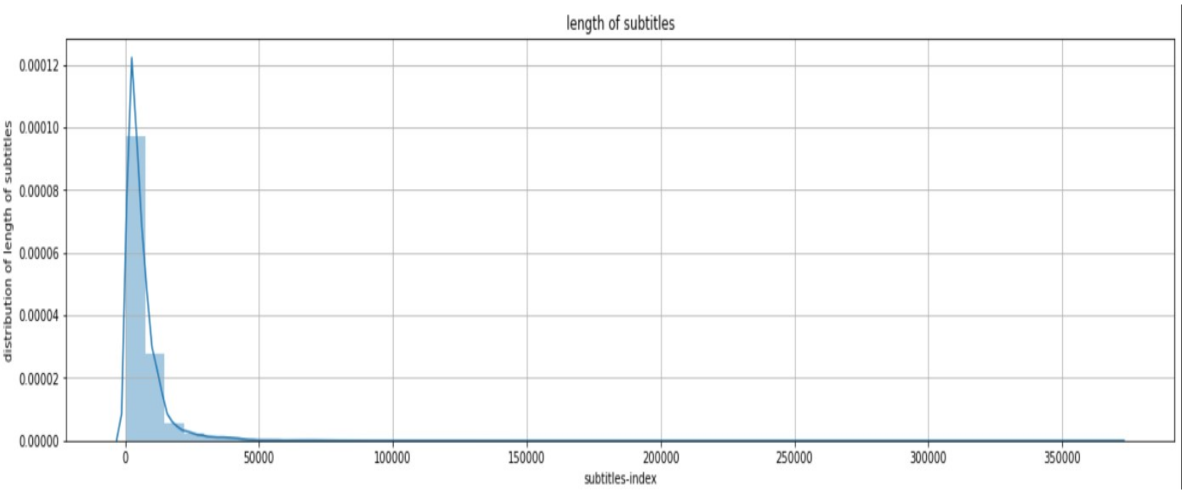


**Figure 2.** Distribution of the length- of -subtitles

### 3.3. Modelling

Different machine learning models were learned based on labelled data belonging to different classes.

The better the quality of training data that is fed into the AI model or quality of ML algorithms, the accuracy of the results increases. The accuracy of model prediction mainly depends on the quality and quantity of training data sets used to train such models.

After analysis, the data was converted into BOW Vector format, and the 270 important unique features/words which were present in at least 5000 documents from the whole 35057 data corpus.

Using Randomized Search in the Scikit learn lib 10 models were trained using k-fold cross- validation.

### 3.4. Machine Learning Models

What follows is a brief explanation of models used in this research.

#### Logistic Regression

Logistic regression is a statistical model that in its basic form applies a logistic function to a weighted linear combination of input variables to model the binary dependent variable.

#### K-NN

K-NN (k-Nearest Neighbor) is a non-parametric classification technique. The basic idea is that you provide a labeled data set, and the algorithm will tell you to which class that unknown data point belongs. The unknown is classified by a simple neighborly vote, where the class of close neighbors “wins”. It’s most popular use is for predictive decision making.

#### Decision-Tree

A decision tree is a decision support tool that uses a tree-like model, whose nodes basically test each example for some conditions on the feature values; based on these, the node branches off into different children; the children in turn provide tests for other features; the leaf nodes provide the class label of test examples. An advantage of the decision tree is easy interpretability by humans.

#### Ridge Classifier

Ridge regression reduces the variance of ordinary least squares regression by imposing a penalty on the size of the coefficients of least squares regression model. The error function to be minimized consists not only of residual sum of squares but also sum of squares of coefficients.

#### GBDT-Classifier

Gradient Tree Boosting or Gradient Boosted Decision Trees (GBDT) is a generalization of boosting to arbitrary differentiable loss functions. GBDT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems.

#### Random – Forest Classifier

Random forest is a technique used in modeling predictions and behavior analysis and generates an ensemble of decision trees, each generated by bootstrapping the data sample. Further, the next feature to be selected for branching in each decision tree is chosen from a subset of the features. Classification is performed by taking a majority vote of the individual classifiers. In this way, random forests reduce the variance of the decision tree model.

#### Bernoulli-NB

Bernoulli-NB implements the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a BernoulliNB instance may binarize its input.

### Multinomial-NB

Multinomial-NB is used for multinomial distribution that normally requires integer feature counts. However, in practice fractional counts such as tf — idf may also work. Gaussian NB.

### Compliment – NB

Complement-NB implements the complement naive Bayes (CNB) algorithm. CNB is an adaptation of the standard multinomial naive Bayes (MNB) algorithm that is particularly suited for imbalanced data sets. Specifically, CNB uses statistics from the complement of each class to compute the model's weights. The inventors of CNB show empirically that the parameter estimates for CNB are more stable than those for MNB.

### Gaussian-NB

Gaussian-NB implements the Gaussian Naive Bayes algorithm for classification. When we are dealing with the continuous data Gaussian-NB is useful. It is used to model data as a Gaussian distribution.

## 4. Performance Evaluation

**Table 1.** Result Table

| Model | F1-Score | Var-F1 |
|---|---|---|
| Logistic-Regression | 0. 889 | 0. 07 |
| KNN | 0. 889 | 0. 08 |
| Decision-Tree | 0. 716 | 0. 05 |
| Ridge-Classifier | 0. 850 | 0. 04 |
| GBDT-Classifier | 0. 867 | 0. 04 |
| RF-Classifier | 0. 853 | 0. 02 |
| Bernoulli-NB | 0. 748 | 0. 03 |
| Mutlinomial-NB | 0. 793 | 0. 04 |
| Compliment-NB | 0. 790 | 0. 01 |
| Gaussian-NB | 0. 467 | 0. 05 |

The results for classification have been shown in Table 1. The top classifiers in terms of F1-score are logistic regression, KNN and GBDT classifier. The more stable classifiers are Compliment-NB, RF classifier and BernouilliNB, based on variance of F1-score. Overall, the RF classifier has the best classification performance and stability with an F1-score of 0.853 and a variance of F1-score of 0.02 respectively.

The base model for Random-Forest Classifier is a Decision Tree and it creates bootstrap samples from main data frame using column and row sampling.

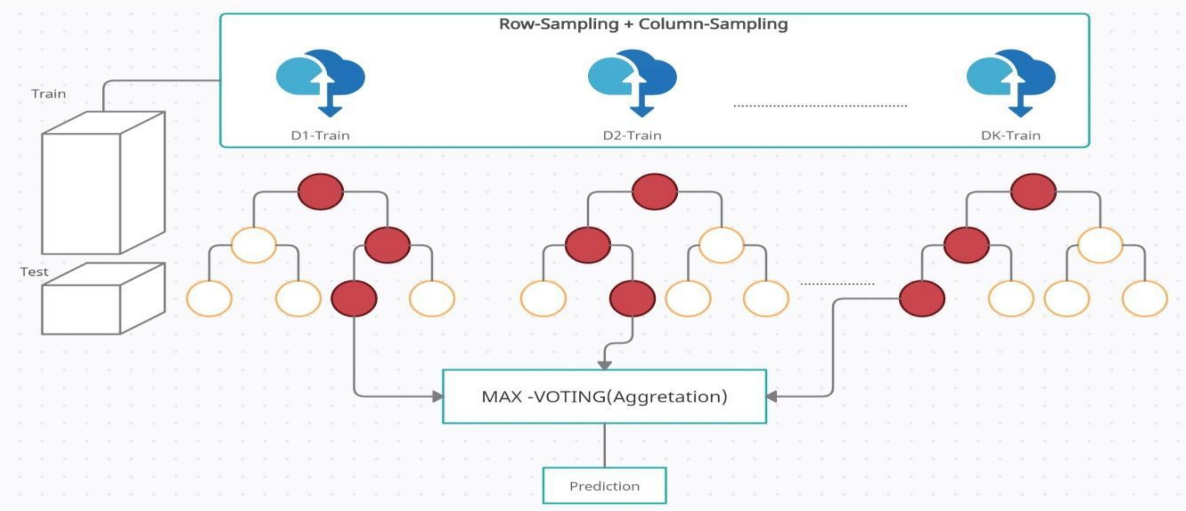


**Figure 3.** Working of Random Forest Classifier

As show in figure - 3, data sample is generated by selecting examples from the data randomly with re-placement. This data sample is used to learn a decision tree. Further, while choosing the next feature for classification, only a subset of randomly selected features is considered. An ensemble model is produced by using all decision trees generated.

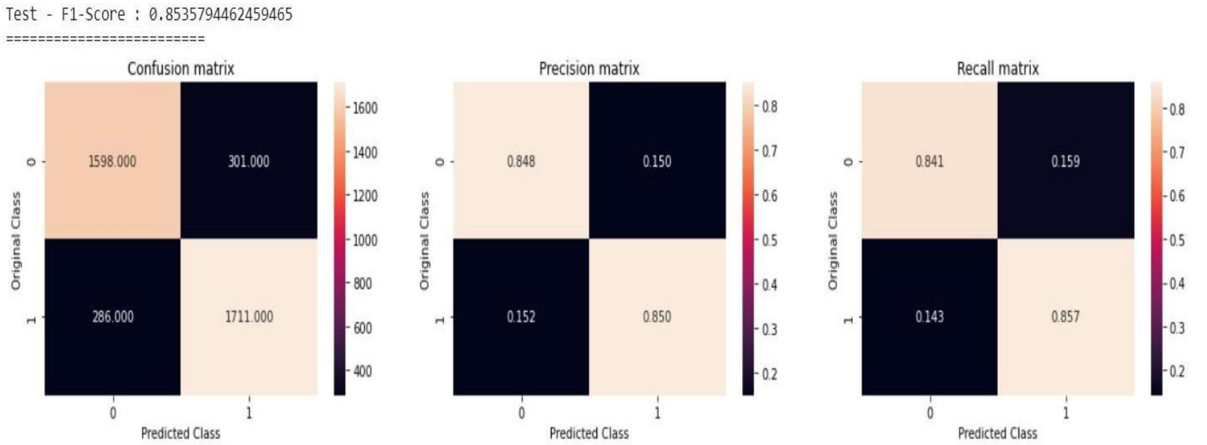
**Advantages of Random-Forest-Classifier**

- Regularization by Randomization: Through row-sampling and column-sampling, Random Forests make K Decision Trees which cuts the variance of the base models during aggregation of the results.

    1. It reduces overfitting in decision trees and helps to improve the accuracy.
    2. It works well with both categorical and continuous values.

- Better Handling of Data The main property of the Random Forest Classifier is its long- term usage - even if the data variance changes over time it will not affect model much because of bagging and sampling of data and different base -models.

    1. It handles missing values present in the data.
    2. Normalizing of data is not required as it uses a rule-based approach.



```
Test - F1-Score : 0.8535794462459465
========================
```

**Figure 4.** Results of Random Forest Classifier

## 5. Conclusion

Text classification has been widely used for many applications, such as news article categorization, social media analysis, and online advertisement. Most text classification techniques represent the text in d--dimensional vector format and apply a machine learning algorithm on the top of it.

Selection of the models were based on two parameters, performance and handling the variance in the data, argument can be drawn of why not using GBDT-Classifier or K-NN. The selection was Random forest classifier as it has good performance while producing a stable model, as indicated by variance of F1-score. It also handles missing values present in the data, and uses aggregation of the base models, which reduces variance in the model and models do not tend to overfit.

In this work, it was demonstrated how videos can be classified by extracting the subtitles, and it was shown that the random forest classifier provides good, stable classification performance. One of the major uses of video classification into academic and entertainment is to filter videos on social networking sites and instant messaging platforms. This serves to increase the productivity of smartphone users who are using social networking and instant messaging platforms throughout the day.

## References

1. Hassan, S., Rafi, M., & Shaikh, M.S. (2011, December). Comparing SVM and Naïve Bayes classifiers for text categorization with Wikitology as knowledge enrichment. In 2011 IEEE 14th *International Multitopic Conference*, (pp. 31-34). IEEE.

2. Kim, S.B., Han, K.S., Rim, H.C., & Myaeng, S.H. (2006). Some effective techniques for naive bayes text classification. *IEEE transactions on knowledge and data engineering, 18*(11), 1457-1466
3. Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. *In Icml,* (Vol. 99, pp. 200-209).
4. Dai, W., Xue, G.R., Yang, Q., & Yu, Y. (2007, July). Transferring naive bayes classifiers for text classification. *In AAAI,* (Vol. 7, pp. 540-545).
5. Mohammad, A. H., Alwada'n, T., & Almomani, O. (2016, August). Arabic Text Categorization Using Support vector machine, Nave Bayes and Neural Network. *In Global Science and Technology Forum Journal of Computing,* (Vol. 5, No. 1, pp. 108-115).
6. Putri, D.A., Kristiyanti, D.A., Indrayuni, E., Nurhadi, A., & Hadinata, D.R. (2020, November). Comparison of Naive Bayes Algorithm and Support Vector Machine using PSO Feature Selection for Sentiment Analysis on E-Wallet Review. *In Journal of Physics: Conference Series,* (Vol. 1641, No. 1, p. 012085). IOP Publishing.
7. Sueno, H.T., Gerardo, B.D., & Medina, R.P. (2020). Multi-class document classification using support vector machine (SVM) based on improved Naïve bayes vectorization technique. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(3)
8. Mekala, D., & Shang, J. (2020, July). Contextualized weak supervision for text classification. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 323-333).
9. Hapsari, D.P., Utoyo, I., & Purnami, S.W. (2020, March). Text Categorization with Fractional Gradient Descent Support Vector Machine. *In Journal of Physics: Conference Series,* (Vol. 1477, No. 2, p. 022038). IOP Publishing.
10. Chantar, H., Mafarja, M., Alsawalqah, H., Heidari, A.A., Aljarah, I., & Faris, H. (2020). Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification. *Neural Computing and Applications, 32*(16), 12201-12220.
11. Alsmadi, I.M., & Gan, K.H. (2020). Short text classification using feature enrichment from credible texts. *International Journal of Web Engineering and Technology, 15*(1), 59-80.