Category Classification of Educational Videos on YouTube through Machine Learning Approaches

Taewon Yoo°, Hyunggu Jung
Department of Software Convergence, Kyung Hee University
xodnjs0208@khu.ac.kr, hgjung@khu.ac.kr

머신러닝을 통한 유튜브 교육 영상 카테고리 분류

유태원°, 정형구
경희대학교 소프트웨어융합학과

Abstract

As YouTube becomes the largest video platform, educational videos in various fields are being uploaded on YouTube. Nevertheless, to deal with the videos that cover multiple topics, viewers and creators of the educational videos still need to classify the video into multiple sub-categories manually. In this paper, we introduce classifiers to categorize of educational videos on YouTube through machine learning approaches. Text in the description of videos was used as a training and test dataset to build the classifiers. We report that the classifiers using text in the description in videos tended to show high accuracy (over 90%). The findings of this study show that educational videos on YouTube can be classified into multiple sub-categories automatically.

1. Introduction

Over the last few years, video media has become a trend's social media platform. As the largest video platform with over 1 billion users [1], YouTube offers videos covering various topics. In particular, YouTube has become a major platform for obtaining information in the educational field. Mirembe et al. has revealed that more than 50% of Ugandan college students say they are watching YouTube videos for educational purposes [2]. Another study has shown that using YouTube videos can increase student engagement, increase critical awareness, and promote deep learning [3]. As the number of students using YouTube as an educational material increases, it becomes more important to identify the right videos efficiently.

YouTube has 15 video categories, such as "Auto & Vehicle", "Beauty & Fashion", "Comedy", "Education", "Science & Technology", and "Music" [4]. One of the constant challenges that learners (e.g., students) face on YouTube is that videos for each category are not automatically subdivided into other sub-categories (e.g., "Science & Technology", "Music", and "Sports"). Thus, the automatic process of classifying educational videos into subcategories may support students in identifying videos covering the topics (e.g., music, science, mathematics) they would be interested in watching.

Previous studies proposed classifiers for category classification and measured their performances [5, 6]. In one study, researchers proposed a method classifying movie review documents into two categories of positive and negative opinions using Support Vector Machine (SVM), Maximum Entropy and score calculation. They evaluated classifiers' accuracy and coverage rate of combination of the methods [5]. Another study showed that poetry can be classified by poet by using classifiers, such as Naive Bayes algorithm, Sequential Minimal Optimization (SMO), C4.5 decision tree, Random Forest and K-Nearest Neighbors [6]. Nonetheless, to our knowledge, little is known about the feasibility of such classifiers in inferring subcategories of educational videos on YouTube. To reduce the gap, we aim to answer the following research questions through the research process (see Figure 1).

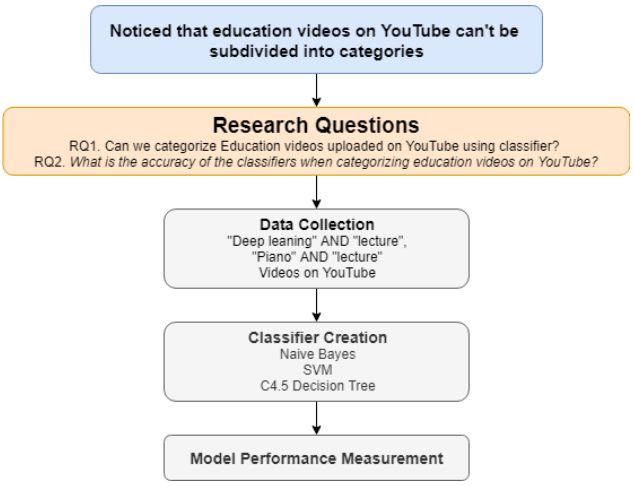

Figure 1. Study Procedure

RQ1: *Can we categorize educational videos uploaded on YouTube using classifiers of machine learning approaches?*

RQ2: *What is the accuracy of the classifiers when categorizing educational videos on YouTube into subcategories?*

* "This research was supported by the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW (2017-0-00093), supervised by the IITP (Institute for Information & communications Technology Planning&Evaluation)."

Table 1. Execution Criteria of YouTube Data API

API execute parameter	Input parameter
part	snippet
order	date
q(query)	1) "Deep Learning" AND Lecture" 2) "Piano" AND "Lesson"
relevance language	en
type	video

In this paper, we aim to categorize educational videos on YouTube by their text in the description section of each video using three popular classifiers based on machine learning approaches: Naïve Bayes, SVM and C4.5 classifiers. We measure the performances of the classifiers by True-Positive (TP) Rate, False-Positive (FP) Rate, recall($\rho = \frac{TP}{TP+FP}$), precision($\pi = \frac{TP}{TP+FN}$), F-measure($\frac{2\rho\pi}{\rho+\pi}$), Matthew's Correlation Coefficient(MCC), Receiver Operating Characteristic (ROC) area and Precision-Recall Curves (PRC) area.

2. Method

The purpose of the classification model is to categorize DL videos as "Science & Technology" and "Piano" video as "Music". To determine if it is possible to categorize educational videos on YouTube into videos with subcategories, we used a classification model to categorize educational videos by their text in the description because it provides viewers with a summary of the video. As a data set, we used the text in the descriptions in the "Deep Learning (DL)" and
"Piano" field. We chose DL as a keyword because DL is considered as one of the topics covered by engineers and scientists. The number of DL articles published in the ACM Digital Library in the 2010s (n=2669) was about 130 times greater than the number of papers published in the 2000s (n=21) [7].

2.1 Data Collection

To get educational videos' text data in the description, we used YouTube Data API and scraped YouTube website. First, we imported all video IDs on YouTube using YouTube Data API v3 at Google APIs Explorer, according to execution criteria (see Table 1) [8, 9]. By executing YouTube Data API v3, we could get total 3607 videos ID, 1106 from DL and 2501 from Piano, in April 2019. Second, we gained the text in the description of each videos by web scraping using Python bs4 module. With these text data, we conducted data selection.

According to Figure 2, 1148 videos were screened at the eligibility stage because they met the following exclusion criteria: 1) text in the description is not written in English, 2) text is duplicated with over 90% text similarity, 3) text is not accessible, and 4) text is missing. As a result, 2469 videos, 610 from DL and 1859 from Piano videos were obtained. Among the obtained videos, we randomly selected 600 videos, one from "Science & Technology"

Table 2. Summary of Datasets from Video of DL and Piano

	(A)	(B)	(C)	(D)
DL	3351	87	2519	99
Piano	442	87	488	107

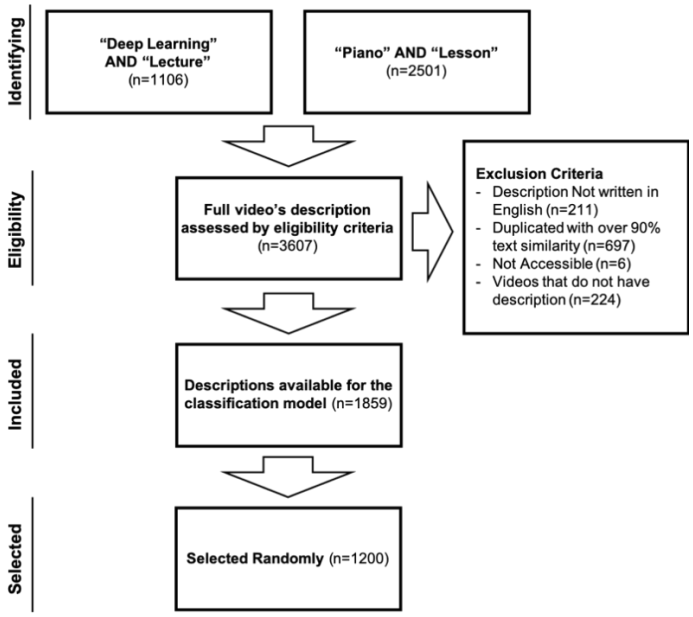

Figure 2. Video Selection Process

and the other from "Music" (i.e., 1200 videos in total) as datasets for creating classifiers.

2.2 Preprocessing

To improve the accuracy of the classification model, we preprocessed the text in the description. First, we removed all the emojis (e.g., "🎶", "❤️", and "🎹") and stop words (e.g., "such", "then", and "so"). We then processed text by stemming and lemmatization. After preprocessing the text, we put labels on each dataset: DL data as "Science & Technology," and Piano data as "Music."

2.3 Creation and Evaluation of Classifiers

We built classifiers using WEKA [10], an open source machine learning software because it offers various classification algorithms. In this study, we used Naive Bayes, SMO-based SVM, C4.5-based decision tree classifiers for categorize educational videos on YouTube. We chose those three classifiers because they were used widely for text classification [5, 6].

2.4 Summary of the Datasets

Table 2 shows the summary of the data set: (A) videos' average of duration (seconds), (B) the average of the number of words in the text in the description, (C) standard deviation of the duration of videos, and (D) the number of words in the description of DL videos and Piano videos. With these data, we built classifiers in two ways. First, we used 70% of these data as training set and the rest of them as a test set. And also did 10-fold cross validation.

Table 3. Summary of Measured Values of Three Classifiers

	70% Training & 30% Test Datasets			10-fold Cross Validation		
	Naive Bayes	**SVM**	**C4.5**	**Naive Bayes**	**SVM**	**C4.5**
Correctly Classified Instance (%)	94.4444	98.8889	94.4444	95.0833	96.8333	93.6667
TP Rate	0.944	0.989	0.944	0.951	0.968	0.937
FP Rate	0.056	0.011	0.056	0.049	0.032	0.063
Precision	0.950	0.989	0.945	0.955	0.968	0.937
Recall	0.944	0.989	0.944	0.951	0.968	0.937
F-Measure	0.944	0.989	0.944	0.951	0.968	0.937
MCC	0.894	0.978	0.890	0.905	0.937	0.874
ROC Area	1	0.989	0.975	0.996	0.968	0.983
PRC Area	1	0.983	0.966	0.996	0.954	0.980

3. Results

We categorized educational videos on YouTube by their text in the description. We found that the educational videos on YouTube can be categorized into subcategories through machine learning approaches. For performance measurement, we measured multiple factors: TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, ROC area and PRC. As shown in Table 3, the results of all 3 different classifiers showed high performance in classifying videos' categories where the accuracy of the classifiers was higher than 90%. SVM showed 98.8% accuracy and was the best among the classifiers.

4. Conclusion

Overall, our research aims to enable learners to search for educational videos efficiently by classifying videos through machine learning approaches. To answer the research questions, we built classification models that categorize educational videos with two topics as an example: DL videos as "Science & Technology" and Piano videos as "Music." We used three difference classification algorithms, Naïve Bayes model, SVM, and C4.5 decision tree. To measure classifiers' performances, we took two tests: One for dividing dataset into 70% for training and 30% for test, and the other for the 10-fold cross validation. Over the 90% classification success rate was taken by all three classification algorithms. The findings of this study showed that educational videos on YouTube can be categorized by their text in the description.

The limitation of this study is that we only considered two categories, "Science & Technology" and "Music" as subcategories of educational videos, out of the 14 categories. The focus of our study was to only evaluate the performances of three known classifiers: Naïve Byes, SVM, and C4.5 for video classification process. In addition, we only used text in the description of each video as a dataset when creating the classifiers. As for future work, we plan on categorizing other educational videos' categories, such as "Sports", "Gaming", and "Food". Moreover, future work still remains to investigate the possibility of other types of machine learning algorithms, such as logistic regression, k-nearest neighbors, and random forest, which were not used in this study. To improve the accuracy of the classifiers, additional metadata such as titles, captions, and comments of videos have potential to be used as training and test datasets.

References

[1] Youtube.com. (2019). Press - YouTube. [online] Available at: https://www.youtube.com/intl/en-GB/yt/about/press/. [Accessed 5 Jun. 2019].

[2] Mirembe, D. P., Lubega, J. T., & Kibukamusoke, M. (2019). Leveraging Social Media in Higher Education: A Case of Universities in Uganda. *European Journal of Open, Distance and E-learning*, *22*(1).

[3] Clifton, A., & Mann, C. (2011). Can YouTube enhance student nurse learning?. *Nurse education today*, *31*(4), 311-313.

[4] Creatoracademy.youtube.com. (2019). [online] Available at:https://creatoracademy.youtube.com/page/lesson/overview-categories [Accessed 6 Jun. 2019].

[5] Tsutsumi, K., Shimada, K., & Endo, T. (2007). Movie review classification based on a multiple classifier. In *Proceedings of the 21st pacific Asia conference on language, information and computation* (pp. 481-488).

[6] Sahin, D. O., Kural, O. E., Kilic, E., & Karabina, A. (2018). A Text Classification Application: Poet Detection from Poetry. *arXiv preprint arXiv:1810.11414*.

[7] Dl.acm.org. (2019). ACM Digital Library. [online] Available at: https://dl.acm.org [Accessed 6 Jun. 2019].

[8] Google Developers. (2019). YouTube Data API | Google Developers. [online] Available at: https://developers.google.com/youtube/v3/ [Accessed 6 Jun. 2019].

[9] Developers.google.com. (2019). Google APIs Explorer. [online] Available at: https://developers.google.com/apis-explorer/#p/ [Accessed 6 Jun. 2019].

[10] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.