Category Classification of Educational Videos on YouTube through Machine Learning Approaches

Taewon Yoo
Department of Software Convergence
Kyung Hee University
xodnjs0208@khu.ac.kr

Hyunggu Jung
Department of Software Convergence
Kyung Hee University
hgjung@khu.ac.kr

INTRODUCTION

- One of the constant challenges that learners face on YouTube is that videos for each category are not automatically subdivided into other subcategories.
- Nonetheless, little is known about the feasibility of classifiers in inferring subcategories of educational videos on YouTube.
- To reduce the gap, we aim to answer the research questions through the research process.

RESEARCH QUESTIONS

- **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?*

METHOD

- The purpose of the classification model is to categorize “Deep learning” videos as “Science & Technology” and “Piano” videos as “Music”, using their text in the description.

DATA COLLECTION

- We collected total 3607 educational video’s text in the description, 1106 from (“deep learning” AND “lecture”) and 2501 from (“piano” AND “lecture”), using YouTube data API according to execution criteria (Table 1).
- Among those videos, we screened and selected videos at the eligibility stage (Figure 1).

Table 1. Execution Criteria of YouTube Data API

API Execute Parameters	Input Parameters
part	snippet
order	date
q(query)	“deep learning” AND “lecture” “piano” AND “lecture”
relevance language	en
type	video

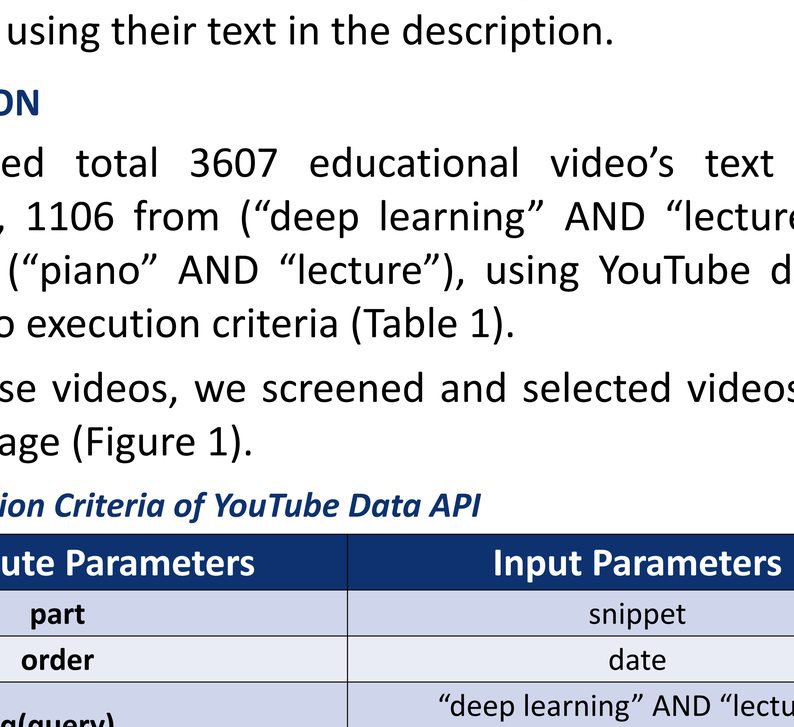
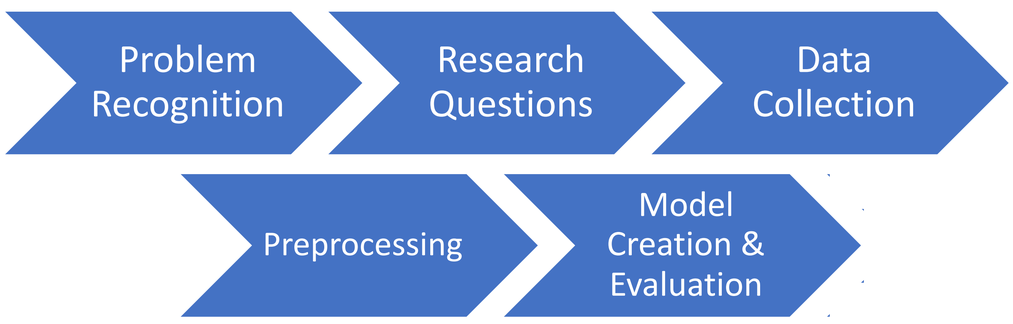


Figure 1. Video Selection Process



Research Procedure

PREPROCESSING

- We removed all the emojis (e.g., “🎶”, “❤️”, and “🎹”) and stop words (e.g., “such”, “then” and “so”).
- We did stemming and lemmatization.

MODEL CREATION

- We built classifiers using open source software, WEKA.
- As a machine learning algorithms, we used Naïve Bayes, SMO-based SVM, C4.5-based Decision Tree.
- We built classifiers in two ways: 1) using dataset as 70% training and 30% test set, and 2) 10-fold cross validation.

RESULTS

Table 2. Summary of Measured Values of Three Classifiers

	70% Training & 30% Test Datasets			10-fold Cross Validation		
	Naïve Bayes	**SVM**	**C4.5**	**Naïve Bayes**	**SVM**	**C4.5**
Correctly Classified Instance (%)	94.444	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

- We found that educational videos on YouTube can be categorized into subcategories through machine learning approaches.
- All three different classifiers showed high performance where the accuracy was higher than 90% (Table 2).

CONCLUSION

- To answer the research questions, we built classifiers that categorize “deep learning” videos as “Science & Technology”, and “piano” videos as ”Music” and measured their performances.
- For future work, we plan on categorizing other educational videos’ on YouTube, and using other types of machine learning algorithms.

KYUNG HEE UNIVERSITY

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).