COMP4801 Final Year Project

HKUCS Graduate Admission Data analysis:

A Multimedia Mining Approach

Final Report

Supervisor: Dr. Reynold Cheng

Group Member: Song Yi Ting (3035124829)

Wang Michelle Yih-chyan (3035124441)

Author: Song Yi Ting

Submission Date: 15th April 2018

Disclaimer

The information published in the report is provided for educational use only. The Department of Computer Science of the University of Hong Kong and the project team is committed to ensuring that the application data and interview videos of the applicants of graduate programs are treated with respect and safeguarded to ensure privacy.

Abstract

In recent years, Educational Data Mining (EDM) has become more involved in the graduate admission process of the Department of Computer Science of the University of Hong Kong (HKUCS). A web application called HKUEDM had been developed in 2017 for extraction, analysis and prediction of text-based data of graduate program applicants in HKUCS. In order to fully utilize the multimedia data, the project aims at enhancing the data mining functionality of the original HKUEDM by exploring interview videos during the admission process. The project created two systems responsible for visual and audio data extraction and analysis. For these applications, several tools have been leveraged, including Affectiva API and Google Speech Recognition API. After the information are extracted, the project analyzes them using multiple data mining techniques. And the project also includes an experiment for evaluating the result prediction function.

Acknowledgement

We would like to express my deep gratitude to Dr. Reynold Cheng, our project supervisor, for his patient guidance and useful critiques for the direction of the project. His willingness to give continuous support throughout the project development has been very much appreciated.

Our grateful thanks also extends to Mr. Jiafeng Hu for his help on collecting admission data as well as providing thorough information on HKUCS graduate program. Without his generous assistance, the project would not have been completed.

We would also like to express our great appreciation to Mr. You Wu and Ms. Fangyuan Xu, who are the group members of Mining HKUCS Graduate Student Data: Extraction, Analysis, and Prediction, for their valuable and constructive suggestions during the planning of the project as well as.

Table of Contents

List of Figures

List of Tables

Abbreviations

Ajax	Asynchronous JavaScript and XML
CSS	Cascading Style Sheets
DOM	Document Object Model
EDM	Educational Data Mining (EDM)
FPR	False Positive Rate
HKUEDM	Educational Data Mining Tool of the University of Hong Kong (i.e. the name of the web tool developed by Wu and Xu [1])
HKUCS	The Computer Science Department of the University of Hong Kong
HTML	Hypertext Markup Language
REST	Representational State Transfer
ROC	Receiver Operating Characteristic
SDK	Software Development Kit
TPR	True Positive Rate

1 Introduction

Each year, the Department of Computer Science of the University of Hong Kong (HKUCS) attracts hundreds of student applicants around the globe for graduate programs. Over the years, the Department has collected numerous admission data regarding student profiles, undergraduate institutions, academic performance and interviewer's comments, which are very valuable in terms of Educational Data Mining (EDM). Educational Data Mining is an emerging discipline which focuses on studying educational-related problems, such as student admission, using data generated from educational settings [2].

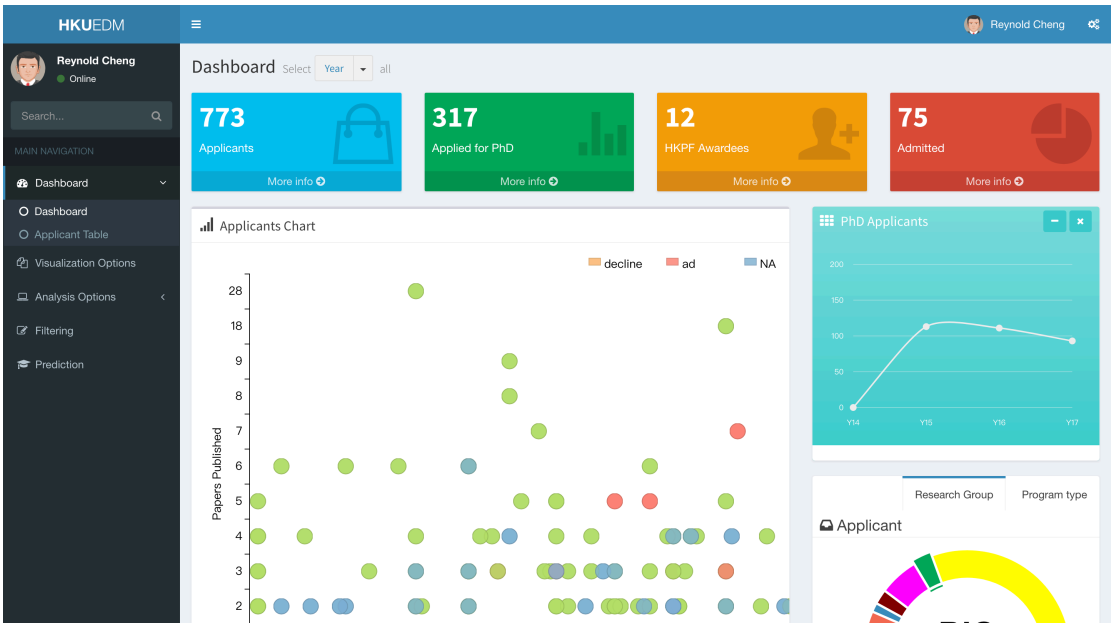

Figure 1 HKUEDM

Last year, HKUCS started to incorporated this new concept into the graduate admission process and started a project called ‘Mining HKUCS Graduate Student Data: Extraction, Analysis and Prediction’. Wu and Xu participated in the project and constructed an educational data mining web application HKUEDM (See Figure 1) [1]. However, the system only makes use of text-based data, such as applicant’s GPA, English examination score and research interests. This limitation excludes a major part of admission procedures – interview. Since HKUCS has recorded admission interviews of all applicants in the past three years (2015, 2016 and 2017), there are actually another area of educational data for exploiting. Therefore, the project will focus on multimedia data mining on HKUCS graduate admission data and provide admission officers with insights into crucial admission information.

The goal of the project is to equip the original HKUEDM with multimedia data analyzing functionality. To achieve this, the new system has to be able to extract useful indicators from interview videos as well as analyze these information to give out prediction. Finally, system integration for the two EDM application is needed in order to present the most comprehensive analysis as well as the best user experience.

The report proceeds as follows. It starts with introducing the project background and outline the project objectives. Then, it proceeds with project scope, including each layer of the system, and the final project deliverable. It later examines the technology used in the project and discusses the implementation methodology in detail. To conclude, it presents the result together with difficulties encountered during the project development and suggests some potential future plans as the next step of the project.

2 Project Background

The basis of the project is HKUEDM, which is a web application for visualizing and analyzing HKUCS graduate admission data. This educational data mining system is created by Wu and Xu last year. After examining the application and having a requirement analysis interview with Dr. Reynold Cheng, the head of graduate admission of HKUCS, and Mr. Jiafeng Hu, several limitations have been identified.

First, the original HKUEDM solely focuses on the analysis of admission data in the initial stage of the admission process, i.e., application submission. Namely, the data processed are all text-based data, such as GPA, undergraduate university, QS ranking

of educational institute and research interests. However, another admission procedure – interview – also plays a crucial role in the evaluation of student applicants, and the corresponding multimedia data generated (interview videos) should be appropriately utilized as well.

Second, multimedia data contains information that cannot be directly handled by traditional data mining models. These information are usually embedded within the data itself, combined with other 'noisy parts' that prevents us from full exploitation. Thus, the interview videos collected by HKUCS have to be pre-processed to extract valuable information before any further analysis is applied.

As a result of the limitations mentioned above, the project will aim at adding the functionality of extracting, analyzing and visualizing of multimedia data to HKUEDM. The detailed objectives will be described in the next chapter.

3 Project Objectives

The goal of this project is to extend the functionality of the original HKUEDM to analyze multimedia admission data, i.e., interview video. Based on the data mining result of multimedia data, the system will be able to provide in-depth analysis of underlying admission strategy as well as predict admission results to help admission officers make accurate and efficient decisions.

Selection Criteria	**Indicators**
Communication Skills	English accent, speaking fluency
Professional Impression	relevant background knowledge, emotion stability
Personal Characteristics	confidence, maturity

Table 1 Interview Selection Criteria

To achieve this objective, the project listed several general interview selection criteria as the evaluation standard for the performance of student candidate (See Table 1). What's more, the multimedia data extraction application should be able to excerpt information indicating whether interviewees are matching these standards. After these information are stored in the database, they should be systematically analyzed using

data mining techniques in order to shed light on the admission decisions of HKUCS. Finally, all of the result should be presented using user-friendly visualization interfaces where admission officers can obtain comprehensive information about all of the applicants.

4 Project Scope

Considering the complexity of the multimedia data in the project as well as the depth of information processing, the project is divided into three layers (See Figure 2).

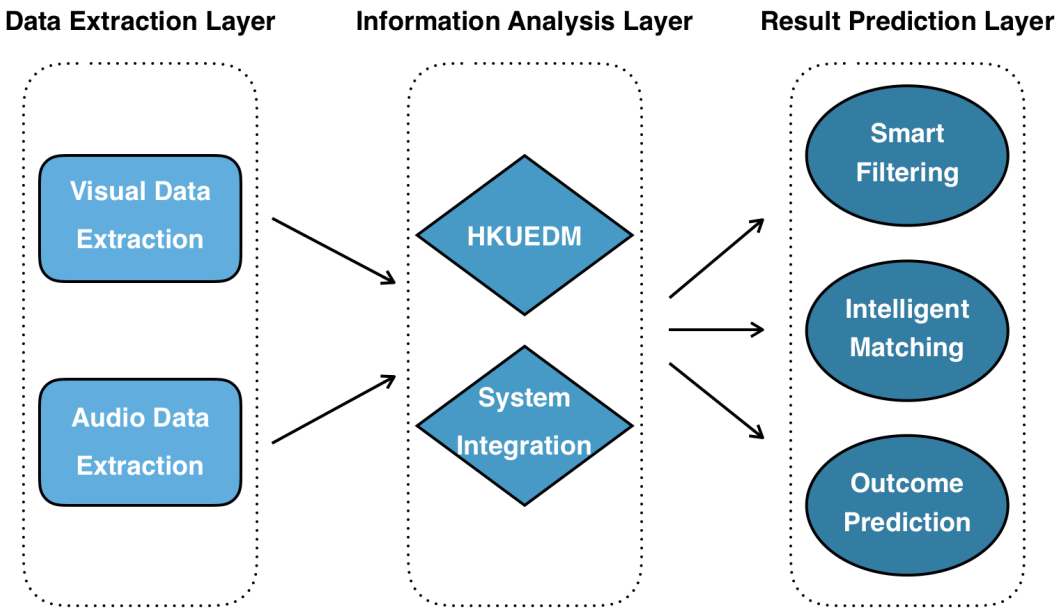

Figure 2 Project Structure

4.1 Data Extraction Layer

The first layer, data extraction layer, focuses on the processing of multimedia data in order to extract useful information. Since the major data source of the project – video – contains information that are nearly inextricably intertwined, it is essential to design methods that are able to extract important messages from them. The two methods are introduced in the following parts.

4.1.1 Visual Multimedia Data Extraction

For visual elements of an interview video, there are two general categories that are commonly viewed as emotion/personality expressing channels – facial expression and body movement (e.g., gesture). Due to time constraints, the project only focuses on the extraction of facial expression and interprets their underlying implication that can be applied to discover the relationship between interview performance and admission results.

4.1.2 Audio Multimedia Data Extraction

As for audio multimedia data, there are more dimensions which can be explored. To start with, the subject who is speaking should be identified in order to distinguish

interviewee from interviews. Later, speech tone, accent, speech fluency and speech context can be obtained for analysis in later stages. These information are expected to provide a more precise estimation of the quality of the interview as well as performance of the interviewee.

4.2 Information Analysis Layer

After the raw multimedia data have been processed, the project will proceed to analyze those information so as to discover hidden patterns of the admission interviews.

4.2.1 HKUEDM

HKUEDM is a web application created by Wu and Xu which provides users with multiple analyzing functions through a set of data mining algorithms [1]. With some adjustments customized for multimedia data, this project will utilize the original system to generate structured classification/rules that may help cast light on the underlying admission strategies.

4.2.2 System Integration

In order to fully exploit the functionalities of HKUEDM, it is crucial to ensure the newly extracted data is compatible with the original design of the system. As a result,

there are several adjustments to be made. First, the integration of MySQL database regarding data consistency as well as data integrity [10]. For each and every record generated from the multimedia data extraction system, it has to be matched with the original application using one or more attributes, such as id number or reference number.

Secondly, the data type of the analyzed result has to be similar to that of original data, i.e., text-based data. This means the newly generated extraction data have to be in numerical form, Boolean form or short character sequence in order to maintain the data consistency as well as functionality of the original analysis model.

4.3 Result Prediction Layer

After both audio and visual multimedia data have been transformed into corresponding factors and analyzed using information analysis system, the last layer – Result Prediction Layer – should be able to suggest suitable applicants to admission officers. This recommendation will be based on the admission patterns found by previous layers as well as the newly added student profile.

5 Project Deliverable

The project deliverable is a collection of applications. The main entrance to the analyzed data will be a web application with data visualization and user-friendly interface.

Figure 3 Visual Data Extraction Application

For visual multimedia pre-processing, the project provides another web application containing out-of-the-box analysis functionality, utilizing Affectiva API which will be introduced in Project Methodology section (See Figure 2) [3].

On the other hand, audio data extraction application will be a collection of functional

systems, each responsible for a component in the data extraction and information analysis layer.

6 Contribution

Name	Responsibility
Song Yi Ting	Visual Multimedia Analysis
	System Integration
Wang Michelle Yih-chyan	Audio Multimedia Analysis
	Result Prediction

Table 2 Project Contribution

7 Project Methodology

This chapter provides a concise explanation of methodology used in the project. It begins with introducing some technologies utilized by the project. It then proceeds with presenting the system structure of the new HKUEDM project. This chapter ends with the detailed implementation of each and every parts of the application.

7.1 Technology Leveraged / Tool Evaluation

This section will give a brief introduction to all of the technology and tools used in the project. The following sections are divided into three parts according to their respective functionality.

7.1.1 Visual Data Processing

To facilitate visual data processing and extract facial expression mentioned in Project Scope, the project chooses Affectiva API as the video/image processor. There are two main reasons why Affectiva is selected. The first and main reason is that it supports multiple practical feature detection, such as emotions (See Figure 4), facial expressions, and appearance. The other reason is that Affectiva supports almost all popular platforms. For example, Linux SDK is constructed in C++ while Unity SDK uses C#. This project selects JavaScript SDK over other languages based on its compatibility with web application since the it is a part of the (new) HKUEDM system.

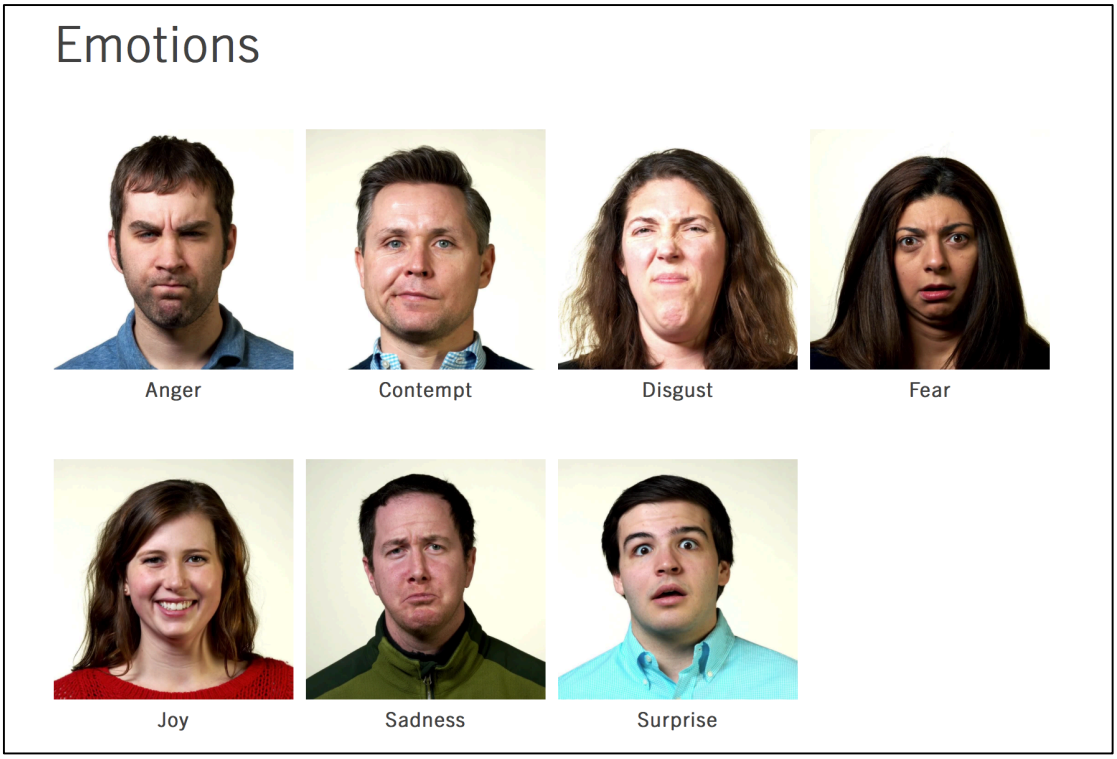

Figure 4 Affectiva API (Emotions)

For front-end implementation, the project chooses the typical web application setup with HTML, CSS and JavaScript. HTML is responsible for the main body of the data extraction system while CSS and JavaScript add style and functionalities to it respectively. In addition to plain JavaScript, jQuery is used to facilitate Ajax communication between server and browser [4]. jQuery is open-source and cross-platform JavaScript library which is designed to simplify client-side scripting. It provides handy function calls which allows users to find, select and manipulate Document Object Model (DOM) in a straightforward manner.

On the other hand, Node.js and Express are used to construct the back-end server of the data pre-processing tool [5], [6]. Node.js is a JavaScript run-time environment which allows server-side scripting while Express is a web application framework for Node.js. These two technologies provide a systematic and convenient method for the project to build a simple yet functional web-based system. Moreover, Node.js can run on different types of platforms, including Linux, macOS and Microsoft Windows, which greatly increases the compatibility of the system. For database, the project decides to continue adopting MySQL database, which is used by original HKUEDM, since the newly analyzed results will be stored and utilized by the system together with the existing ones.

7.1.2 Audio Data Processing

For audio data pre-processing, the project also has to extract useful indicators and contents from this complex multimedia data type. As a result, several tools are leveraged to facilitate the processing.

First, Google Speech Recognition API is selected for speech-to-text extraction [7]. It is a technology converting audio to text using powerful neural network models. It is pre-

trained using machine learning techniques and supports real-time transcript return. All of the characteristics above combined make it stand out from other tools available.

After the speech contents are retrieved, the project further breaks down the result into smaller units and turn them into numerical expressions for exploring underlying patterns. Below is a list of methodologies used:

Bag-of-words

Bag-of-words is a model for simplifying natural language. In the model, a text, such as a sentence or a document, is represented as a bag full of words. However, there is no grammar or ordering inside each bag. The only feature is the number of times each word appears in the text.

For example, a sentence '*She is very happy to be here. He is happy, too*' will be turned into `BoW = {"She":1, "is":2, "very":1, "happy":2, "to": 1, "be": 1, "here": 1, "He": 1, "too":1}`.

Bigram bag-of-words

Bigram bag-of-words is an extension of bag-of-words with takes 2 contiguous words as a unit of item. For example, the sentence ‘dog that barks does not bite’ has a bigram of *dog that*, *that barks*, *barks does*, *does not*, *not bite*. Any further procedure will be based on this bigram.

Tf-idf

Tf-idf stands for term frequency and inverse document frequency. It is designed for determining the significance of a word in a collection of documents and corpus. Term frequency is the number of times a term appears in a document while inverse document frequency is the inverse of the proportion of documents that contain a certain term. We combine these two terms using the formula below:

$$tfidf(t, d, D) = tf(t, d) \cdot idf(t, D)$$

As the term appears more in a document, its term frequency goes up, which implies its importance in the particular document. Nonetheless, for swords like ‘is’, ‘the’, ‘to’ that are common in most documents, their significance will be overestimated. Therefore, inverse document frequency is incorporated to diminish the weight of

terms with high frequency but low relevance.

Word2Vec

Word2Vec is a group of models trained using shallow neural network. It takes texts as input and output a vector space of several hundred dimensions where each word corresponds to a vector. Words that bear similar meanings will be located in close proximity.

Doc2Vec

Doc2Vec is an adaptation of Word2Vec. It tries to create a numeric representation for a document using almost the same method as Word2Vec. However, it adds a small extension to the model called document vector and trains it with other word vectors. After process have completed, the document vector can be seen as the concept of the document.

7.1.3 Original HKUEDM System

Since the project leverages the original HKUEDM as analysis platform, it contains the entire web application created using Python as development programming language [8]. Wu and Xu adopted Django as web framework due to its characteristics of reusability and scalability [9]. They also chose d3.js, a JavaScript library that manipulated Document Object Model (DOM) through user-friendly APIs, for data visualization [11]. As for database, the project continues to use MySQL database from the original system because of its vast user base as well as being free and open source [10]. Being large groups of developers' choice makes it easy to find supports on public forums when encountering difficulties while using open source software reduces the cost of development.

7.2 System Structure

This sections illustrates the architecture of the new HKUEDM system. It first introduces the general framework of both front-end and back-end of the complex application. Then, the second part focuses on the design of MySQL database classes.

7.2.1 Overall Framework

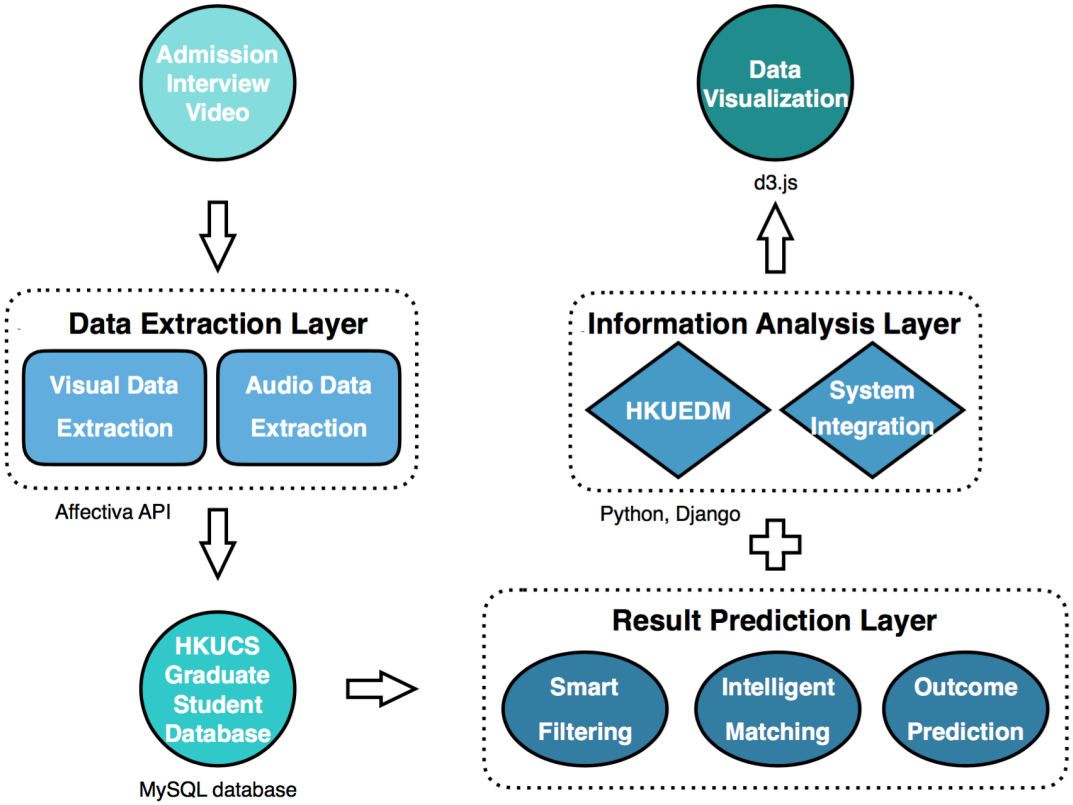

Figure 5 Overall Project Framework

Figure 5 shows the detailed structure of the project. The data source of the project is the admission interview videos in previous years (located in the upper-left corner). Through the first layer – Data Extraction Layer, they will be processed using both visual and audio data extraction applications. The raw video will be turned into several forms of functional data type and stored in HKUCS Graduate Student Database under the database schema shown in the following section.

After all the data are transformed, the project will utilize the new HKUEDM, which is customized for both parameterized/text-based data and multimedia data, to analyze these information. The Information Analysis Layer will apply several popular data mining techniques on the newly processed multimedia data so as to discover informative admission patterns needed in result prediction. Then, Result Prediction Layer will make use of the patterns found to support a recommendation system with Logistic Regression.

Finally, the result will be visualized using ring diagram and customizable bubble chart to give the users a clearer glimpse of the analyzed admission data. Also, users are able

to select different sets of attributes they are interested in and compare applicants from different academic years based on these attributes.

7.2.2 Database Design

Figure 6 MySQL database class diagram

Figure 6 illustrates the structure of MySQL database, which consists of two major parts – *applicants* and *facial_expression*. The *applicant* table contains the original text-based data extracted by Wu and Xu. It represents the applicants' background information regarding their academic performance (e.g., *ug_gpa*, *toefl*) and educational institute attended (e.g., *QSRanking*) as well as admission-related records (e.g., *ad_round*, *ad_result*).

On the other hand, *facial_expression* class is used for storing all the visual multimedia data (i.e., facial expression) in the project. *TypeID*, *emotionID* and *expressionID* column in *facial_expression* table each relates to one corresponding table for detailed information about the type, emotion and expression the interviewee expressed at the moment of interview (represented by *second* column). The *emotion* and *expression* table represents the information extracted from Affectiva API resprectively while the *type* table is indicating the situation the student interviewee is in, e.g, default, self-introduction, listening to questions or answering questions.

7.3 Implementation

This part of the report will describe the implementation methods of each component of the project in detail. The following sections will introduce these components in accordance with the layer it belongs to.

7.3.1 Data Extraction

As discussed in Project Background chapter, due to the fact that multimedia data usually contains information that are inextricably intertwined, data extraction plays a critical part in the project, especially for information analysis. Therefore, the sections below will discuss the methodology of this layer from the beginning (data collection) to the end (data pre-processing).

Data collection

The project collected admission data of 773 HKUCS graduate program applicants from previous four years (2014, 2015, 2016 and 2017) so as to facilitate integration with the original HKUEDM system. In addition, the project also gathered a total of 208 interview videos from 2015, 2016 and 2017 for multimedia analysis.

Data Pre-processing – Visual

After collecting all of the interview videos available, the project constructs a visual multimedia extraction tool that is able to obtain informative factors about HKUCS graduate program applicants through interview videos. As mentioned in Project Objective chapter, the information extracted have to be able to reflect the performance of the interviewee. Therefore, the project evaluated several existing facial expression recognition tools available and selected Affectiva.

Referring back to the Technology Leveraged section, Affectiva has great compatibility with multiple platforms and provides developers with powerful functionalities. The project chooses Affectiva JavaScript SDK for the front-end data processing to go along with the back-end web server created using Node.js and Express. For front-end web page, the project first created a HTML file as the main body for embedding interview videos. Then, a CSS file is added to the main body for styling as well as immediate visualization of data processing. Last but not least, a JavaScript file is created for data extraction. For every second, the screenshot of the video will be extracted as a canvas object and processed through Affectiva APIs.

Then, Affectiva will return an array of faces detected. Each ‘face’ object has a series of key-value pairs indicating the score of each attribute respectively (See Table 3).

Emotions			
Anger	Contempt	Disgust	Fear
Joy	Sadness	Surprise	
Facial Expressions			
Attention	Brow Furrow	Brow Raise	Cheek Raise
Chin Raise	Dimpler	Eye Closure	Eye Widen
Inner Brow Raise	Jaw Drop	Lid Tighten	Lip Corner Depressor
Lip Press	Lip Pucker	Lip Stretch	Lip Suck
Mouth Open	Nose Wrinkle	Smile	Smirk
Upper Lip Raise			
Appearance			
Age	Ethnicity	Gender	Glasses

Table 3 Analyzed result of Affectiva API

After a certain amount of analyzed results have been collected, it will trigger an Ajax call using jQuery API to send these results to the back-end server for storage. The back-end server of this visual data extraction system follows Express structure where a number of ‘routers’ are created to handle different RESTful requests. Among them, ‘/store’ router is the most frequently used. It is responsible for storing data transferred trough HTTP Ajax call from the front-end JavaScript file. When a request comes in, the’/store’ router first connects to MySQL database and turn the data in the request body into valid SQL statements for insertion. It then sends out corresponding SQL commands to the database and inform the front-end whether the operation is successfully completed.

Data Pre-processing – Audio

For audio data pre-processing, the project first extracts the audio parts from the recorded interview videos and makes minor adjustments to increase the volume and quality for further processing. Then, speech contents are retrieved from audio files using Google Speech Recognition API. For students with multiple interviews, the contents are combined together using applicant id.

Simply obtaining the speech text does not satisfy the Project Objective of multimedia data analysis. The project further transforms these contents into numerical format in order to discover more practical indicators for result prediction. The methods used include bag-of-words, bigram bag-of-words, tf-idf, word2vec and doec2vec introduced in Technology Leveraged chapter. These methods not only distinguish the importance (weighting) and relevance of terms in the interview texts but also virtualize the concept of terms and documents. Finally, these results are stored in the database together with admission result of each student applicant for later analysis.

7.3.2 Information Analysis

This section introduces implementation methods of components in the Information Analysis Layer. Since HKUEDM is implemented by Wu and Xu in the 'Mining HKUCS Graduate Student Data: Extraction, Analysis and Prediction' project, the following part will only discuss system integration of the project.

System Integration

After the multimedia extraction, the results are of different forms as that in the original HKUCS Graduate Student Database. Thus, the first and most important part

of the system integration is to convert the newly processed data into data types that is compatible with the original HKUEDM. Take visual multimedia for example. The data extracted are based on timestamps in the video plus the video title as the identifier. Thus, the project needs to aggregate all of the results regarding one applicant and calculate an index to present his or her overall performance in the interview(s).

Another adjustment lies on data mining models. Due to the difference in nature between text-based data and multimedia data, there may be significant inconsistency after combining the extracted results directly. As a result, the project has adjusted some parameters in the data mining model in order to support the best result prediction functionality.

7.3.3 Result Prediction

Due to the difference in text-based and multimedia data extraction, the project designs a new information processing and result prediction model adapting from the original HKUEDM data mining models. The model consists of Random Forest Classifier and Random Forest Regressor in Scikit-learn for investigation into multimedia results [12].

The project divides analyzed data in the database into training set and prediction set at the ratio of 3 to 1. For training set, it is used as training data for the data mining model mentioned above. After the system is trained, the prediction set is used as the input for evaluation of the system prediction accuracy. In addition to default settings of the data mining model, there are some parameters available for adjustments in order to provide better prediction results. The project also repeatedly trains and tests the extracted multimedia data and adjust the model parameters accordingly to construct the best prediction functionality possible.

8 Experiments, Results and Implication

In order to examine the correctness of the prediction and the best combination of numerical expression and data mining model, the project conducted an experiment by cross-pairing all possible combinations. Additionally, since regression and classification produces different types of results, the comparisons and evaluation below will be based on the same data mining technique.

To evaluate the result of each combination, a statistical technique called 'Receiver Operating Characteristic' (ROC) is required [13]. The ROC curve comprises two components – true positive rate (TPR) and false positive rate (FPR), each representing y- and x-axis respectively. True positive means a prediction and the actual result are both positive while false positive means positive prediction plus negative actual result. Therefore, the ROC graph is also sensitivity versus (1 – specificity) plot, which means the curve across from the left-bottom to right-upper corner of the plot serves as the threshold for true/false positive and negative rates. This also implies that the larger of the area in the upper-right corner of the ROC graph, the better the prediction combination is. The paragraphs below will discuss the best and worst results in the

experiment and raise some implication from the results mentioned.

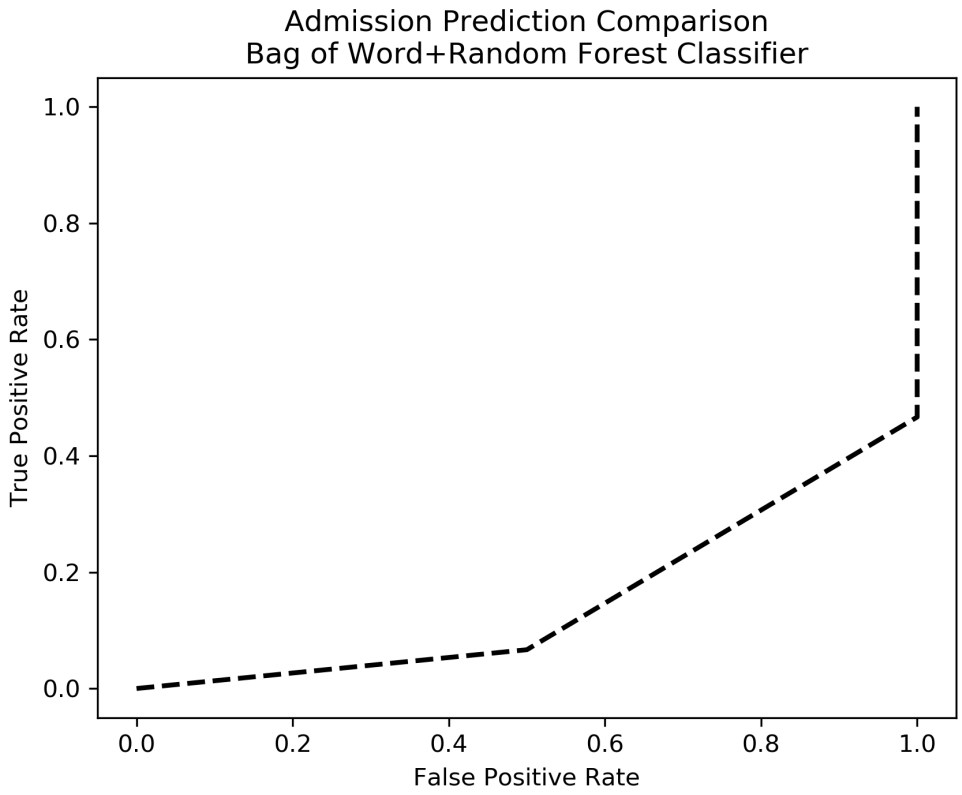

Figure 7 ROC graph for bag-of-word and random forest classifier

For random forest classifier, the best result is obtained from bag-of-word (See Figure 7). And the worst case comes from the bigram bag-of-word (See Figure 8). The differences may result from the proportion of common words in the documents. Since the speech contents are not able to form a large dataset, the effect of the common words in the bigram bag-of-word may not be effective enough in the analysis, causing the prediction result to be only as good as random guessing.

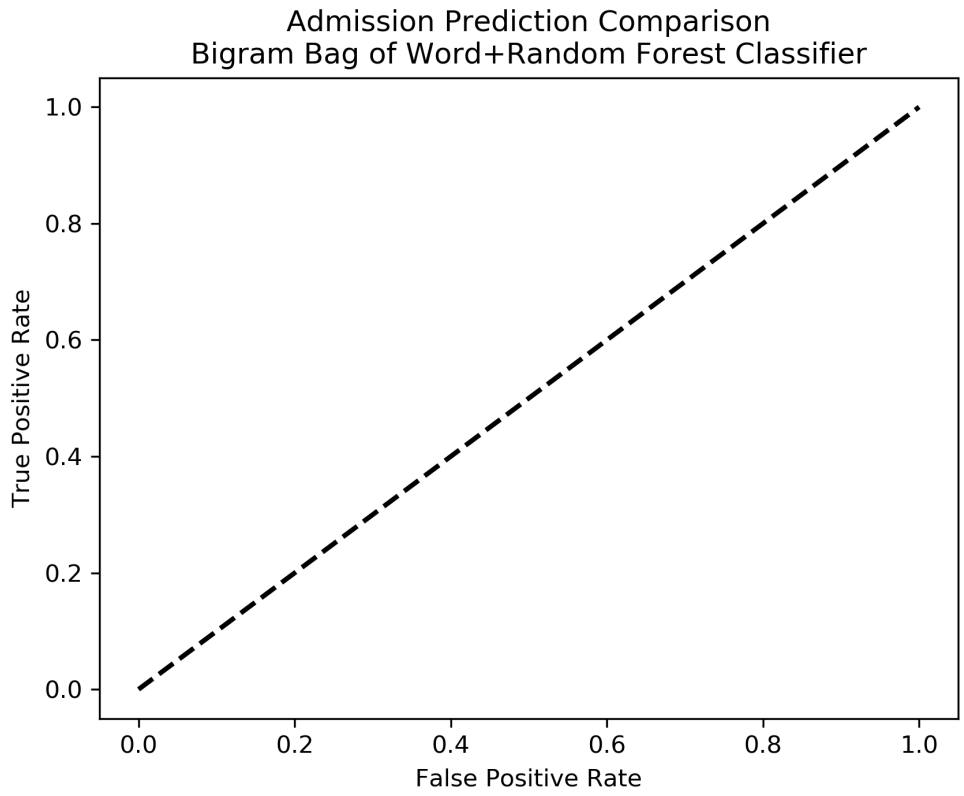

Figure 8 ROC graph for bigram bag-of-word and random forest classifier

For random forest regressor, word2vec yields the best result (See Figure 9) while bag-of-word results in the worst outcome (See Figure 10). This contrast may be caused by the relevance of words in the content analyzed because word2vec preserves the relationship and meaning between each word while bag-of-work only record the occurrence of words in a document.

9 Difficulties and Proposed Solutions

In the course of project development, this project had encountered several difficulties. The first difficulty is the inconsistency of quality of interview videos. Since the source of the admission data comes from three academic years (2015, 2016 and 2017), the camera settings and background environments of the interview vary from one video to another. If the quality of the video is too low or the interviewee is too distant from the camera to a certain extent, Affectiva APIs might not be able to properly process the images extracted. For example, “round_5_id_179_videoId_36_MVI_0097” and “round_5_id_179_videoId_55_MVI_0148” are interview videos of the same applicant in the same meeting room. Nevertheless, the interviewee sat more distant away from the camera in the latter video, which makes her face hard to be detected and, thus, decrease the number of facial expression index extracted. This may lead to data inconsistency and in turn affect the result of data mining.

Despite the fact that it may be time-consuming, a possible solution is to re-process the interview videos in past years to zoom in the interviewee in the frame. Also, for future admission interviews, it is important to make the quality of filming more consistent to

facilitate any further data processing.

Another difficulty is the inconsistency between application data and interview video title. In the original HKUCS Graduate Student Database, each student application record is identified by *idnum* and *reference_no* columns. Nevertheless, interview video title does not always contain a reference number as the application. What's more, the names and university attended in the *applicants* table usually do not match with the content in the interview (self-introduction). This data integrity issue may come from data contamination between transfer of systems or MySQL database version upgrade.

To solve it, a potential solution is to export the data in MySQL database in different forms that are more static, e.g., JSON or XML. When developers in the future tries to access the database, they will be able to cross-check between multiple versions of the data and decrease the chance of data contamination.

10 Conclusion and Future Works

In order to fully exploit the admission data stored in HKUCS Graduate Student Database, the project aims at equipping the original HKUEDM with multimedia data analysis functionality. A visual and an audio multimedia data extraction application have been constructed for the Data Extraction Layer. In addition, the information extracted have also been analyzed and the result prediction function is also evaluated.

An important development area of the project is to discover the full potential of the multimedia information extracted, such as proper interview section categorization and natural language processing on admission interview. By further exploration, we hope that the HKUEDM system can be adopted by other faculties in the University to assist more admission officers to make correct and efficient decisions.

References

[1] Wu, Y, Xu, F. Mining HKUCS Graduate Student Data: Extraction, Analysis, and Prediction. Hong Kong (PRC): The University of Hong Kong; 2017. Available at: http://www.cs.hku.hk/programme/projects/csfyp/csfyp.jsp/.

[2] Romero C, Ventura S. Educational Data Mining: A Review of the State of the Art. IEEE Trans. Syst., Man, Cybern. C. 2010; 40: 601-618.

[3] Affectiva [Internet]. Emotion Recognition Software and Analysis [cited 2018 Apr 14]. Available from: https://www.affectiva.com/.

[4] jQuery [Internet]. jQuery write less, do more [cited 2018 Apr 14]. Available from: https://jquery.com/.

[5] Node.js [Internet]. Node.js [cited 2013 Apr 14]. Available from: https://nodejs.org/en/.

[6] Express [Internet]. Express – Node.js web application framework [cited 2018 Apr 14]. Available from: http://expressjs.com/.

[7] Google Speech Recognition API [Internet]. Cloud Speech-to-Text – Speech Recognition | Google Cloud [cited 2018 Apr 15]. Available from: https://cloud.google.com/speech-to-text/.

[8] Python [Internet]. Welcome to Python.org [cited 2018 Apr 14]. Available from: https://www.python.org/.

[9] Django [Internet]. The Web framework for perfectionists and deadlines | Django. [cited 2018 Apr 14]. Available from: https://www.djangoproject.com/start/overview/.

[10] MySQL [Internet]. The world's most popular open source database [cited 2018 Apr 14]. Available from: https://www.mysql.com/.

[11] D3.js - Data-Driven Documents [Internet]. D3.js - Data-Driven Documents. [cited 2018 Apr 14]. Available from: https://d3js.org/.

[12] Scikit-learn [Internet]. Scikit-learn: machine learning in Python — 0.19.1 documentation. [cited 2018 Apr 15]. Available from: http://scikit-learn.org/stable/index.html.

[13] Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Diagnostic Radiology. 1982; 143.1: 29-36.