Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Haoqi Gu April 1st

Level-based Resume Classification on Nursing Job Positions

By

Haoqi Gu

Jinho D. Choi

Advisor

Department of Mathematics

Jinho D. Choi

Advisor

Yuanzhe Xi

Committee Member

Bree Ettinger

Committee Member

2020

Level-based Resume Classification on Nursing Job Positions

By

Haoqi Gu

Jinho D. Choi

Advisor

An abstract of

a thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Department of Mathematics

2020

Abstract

Level-based Resume Classification on Nursing Job Positions

By Haoqi Gu

In this thesis, we mainly focus on documents of real application resumes. Different from most similar works, we are not categorizing resumes into the suitable groups, for example, IT job resume, medical care job resume, teachers resume, and so on, but we will categorize application resumes on a specific level-based job position called Clinical Research Coordinator from the School of Nursing at Emory University. The job position has 4 different levels, CRC I, II, III, and IV, for applicants to apply to and we aim to write an algorithm to classify resumes into these 4 levels based on their content. Methods used are string matching, feature vectors, bags of words and ensemble models. The best model to predict the admission result of a resume reaches 66.89%.

Level-based Resume Classification on Nursing Job Positions

By

Haoqi Gu

Jinho D. Choi

Advisor

A thesis submitted to the Faculty of Emory College of Arts and Sciences

of Emory University in partial fulfillment

of the requirements of the degree of

Bachelor of Science with Honors

Department of Mathematics

2020

Acknowledgements

Firstly, I would like to thank my advisor, Dr. Jinho D. Choi, for giving me such a valuable opportunity to do research with him. He is a very famous professor at Emory University and many students ever wanted to work with him. I still remembered my experience in his class last year and his expertise and teaching style impressed me deeply. He always gave me great help and insightful suggestions when I felt frustrated. He led me into the NLP world and I really learned a lot since last summer. He is the teacher who made me realize what research is and doing research feels like. He means a lot to me and he is truly the first advisor in my academic life.

Secondly, I want to thank Drs. Yuanzhe Xi and Bree Ettinger for attending my honor thesis defense as committee members. They are great professors who once helped and influenced me a lot in mathematics on numerical analysis and probabilities.

Finally, I want to thank my parents for supporting me study abroad in the U.S so that I could have more opportunities to get in touch with great professors in any academic field. They also supported me mentally during my hard time when I met difficulties and doubted myself. They respected my choices and trusted me and gave me confidence.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Document classification is a scientific method to categorize documents into several groups so that documents in one group share similar characteristics. This method has been widely used in classifying texts, images, music, etc. Within these practical areas, text classification has been applied on many text datasets. Among these text datasets, however, resume dataset is the field that few researchers pay attention to. In this thesis, the main task is to analyze how text classification perform in classifying job application resumes. As we know, many resumes may share same formats. For example, most resumes contain the information about education, work experience, or activities. In addition, resumes can be evaluated qualified or not if certain job requirements are given, so it is possible that resumes can be classified to different qualities or different levels based on the information they provide. Therefore, information about whether qualified resumes share same good

attributes or whether unqualified resumes have deficient content on certain information is a promising field that can be done some research so that certain patterns can be found in all the resumes. For dataset, resumes from job application are enclosed in companies and it is difficult to get a big dataset from various sources. Therefore, in this theis, a new small real dataset has been created in order to do this research. This dataset of resumes is created by the School of Nursing at Emory University. In this thesis, facts about the dataset and methods used such as rule-based classification and unsupervised classification will be introduced in detail later.

In Chapter 2, related work on classifying resumes will be exclusively listed and introduced. Comparison between the goal of their work and this thesis will present more difference from each other. Uniqueness of the dataset used in this thesis will be presented in order to show the distinctness of this thesis. Chapter 3 mainly give the new dataset statistics. Since this dataset is newly created, the annotation process will also be introduced. Afterwards, the preprocessing progress will be explained in detail.

Chapter 4 states all the approaches that are used in this thesis. The approaches used are some direct models like string matching, or some machine learning models based on feature vectors and bags of words as well as some ensembel

models.

Chapter 5 will combine all the experiment results from chapter 4 and do further comparison and analysis. In addition, since the specialty of this new dataset, reliability of the result will be analyzed through error analysis.

Finally, self-reflection on this research will be wrapped up and future possible work that can be done on this dataset or future improvement on the approaches will also be suggested for anyone who shares the same interest on resume classification.

Chapter 2

Background

In this part, related works about resume classification will be discussed (Section 2.1). Compared with the related works, difference (Section 2.2) between their goal of resume classification and our goal of resume classification will be listed. Uniqueness of this dataset and more advantages of analyzing this dataset (Section 2.3) will also be mentioned.

2.1 Resume Classification (Job Category Classification)

Previous examples of resume classifications can be concluded as a classification more like putting each resume into the domain of job category it belongs to and it is suitable to apply to. Because resume data is costly, sensitive and thus hard to achieve, one work used domain adaptation on a number of available job description snippets and indirectly classified resume data

of job applicants into 27 different job categories by using convoluted neural networks. [13] Another work classified the resumes into a suitable domain based on the applicant's interest, education, work experience, expertise, etc. mentioned in the resume.[7]

2.2 Resume Classification (Level-based Position Classification)

Compared with previous works, this thesis mainly introduce and explain a brand new way of classifying resumes in a more practical case. In real world, deciding whether the resume is overqualified, suitable or not qualified is always a down-to-earth and tedious process which requires much labor force from admission. In this thesis, the goal is to learn from a new dataset, which will immediately explain in Section 2.3, and classify application resumes from one specific job position into different expertise levels based on their contents, not classifying into different job categories as previous works.

2.3 Dataset Uniqueness

The dataset used in this thesis has been recently collected from the real resume data from the School of Nursing at Emory University. The School of Nursing

at Emory University posts many positions for applicants every year and every job description illustrated clearly the position requirements. Within those jobs, the Clinical Research Coordinator is the most popular one and looking for most number of employment. In addition, the Clinical Research Coordinator is divided into four levels of positions: level I to IV. Each applicant's resume clearly states the level it wants to apply to; however, the result decided by the admission can always be different from the applicants' self-positioning. Therefore, this dataset models a real life admission process: what applicant resume a certain position can receive and what suitable position level the resume should be admitted to. This kind of dataset is rare online and few people did any similar research before.

Chapter 3

Dataset

In this chapter, the new dataset will be introduced in detail, including the metadata and important statistics result (Section3.1) and we will discuss how it is useful and effective when doing resume classification. Annotation process and rules will be explained to state the quality and accuracy of the dataset(Section 3.2). Due to the specialty that the dataset is brand new, preprocesses are highly necessary to be mentioned in order to clarify the importance and reliability of doing so(Section 3.3, 3.4, and 3.5). The preprocesses mainly includes file format conversion (Section 3.3 and 3.5), section extraction (Section 3.4 and 3.6), manually extraction accuracy check (Section 3.7), and preprocessed data introduction (Section 3.8).

3.1 Dataset Statistics

The new data set was created based on the 6512 application resume pool from the School of Nursing at Emory University. All the application resumes here applied the specific job, Clinical Research Coordinator, which was divided into four different levels: CRC I, CRC II, CRC III, CRC IV. In addition, for each level of CRC position, it may have multiple different CRC jobs. For example, more than one CRC job may have the same CRC level. Therefore, there are 108 jobs for CRC I, 88 jobs for CRC II, 29 jobs for CRC III and 6 jobs for CRC IV. Out of the 6512 unique resumes, in other words, out of the 6512 applicants, some of them may apply multiple jobs in the same level or across the levels, so there were totally 25027 applications. Due to multiple applications from one applicant, one more necessary cleaning process was to divide applicants into groups by their highest will of application. For example, if one applicant both applied for CRC I and CRC II, he or she should be grouped into CRC II applicant by his or her highest level applied. In this way, the ratio of applicants in the four levels was 28:12:8:2. Then, 2025 resumes were randomly selected to form the dataset: 1134 CRC I applicants' resumes, 486 CRC II applicants' resumes, 324 CRC III applicants' resumes and 81 CRC IV applicants' resumes. Annotation of these 2025 resumes will

be explained more in detail in Section 3.2.

3.2 Annotation

Annotation of these 2025 resumes was done by Dr. Elaine Fisher, Rebecca Thomas, Charlie Williams and Sabrina Sabir from the School of Nursing. They are all experts and responsible for hiring new workers in the School of Nursing. Additionally, among 2025 resumes, 250 resumes got double checked by more than one person. Since they are all experts and responsible for hiring new workers, the annotation done by them is real, professional and trustable. In the annotation, each resume has been labeled whether admitted or not to what level of CRC job, which means there are five categories after annotation: Not Admitted, CRC I, CRC II, CRC III, and CRC IV. Along with the annotation, they also cleaned up a guideline file including all the rules and standards when deciding or judging the admission result of a certain applicant. This guideline was used in the rule-based method to make string match for classification and this method will be introduced in detail later(Section 4.1)

3.3 Format Conversion (DOCX)

The two formats that the 2025 resumes have are PDF and docx. Conversion and extraction are separated for PDF (Section 3.5 and 3.6) and DOCX (Section 3.3 and 3.4) files.

For DOCX files, as we want to keep the writing layouts in the resume so that we can extract different section effectively, conversion DOCX resumes into HTML format is ideal. HTML stands for HyperText markup Language and the most distinct attribute is that it has HTML tags for texts to represent the structure of the texts. We used the package Pypandoc to make this conversion happen. One example will be given in Section 3.4 to illustrate why HTML is chosen and how it performed when extracting sections.

3.4 Section Extraction (DOCX)

From one of the resumes, a section about education in this resume can be converted into a HTML format like this.

$<h1$ id="education"$>$ EDUCATION$</h1>$

$<p>KellerGraduateSchoolofManagement</p>$

$<p>SummaCumLaude</p>$ 3.94

$/p >$

$< p >< strong > June 2017 < /strong >< /p >$

$< p > Masters of Public Business Administration < /p >$

. . . . . .

$< h1$ id=”$employment-history$” $> EMPLOYMENT HISTORY < /h1 >$

We can see from the above picture that section titles like EDUCATION, EMPLOYMENT HISTORY shared the same HTML tag, ¡h1. . . ¿ in the beginning and ¡/h1¿ in the end. As a result, for each resume, we firstly convert its form to HTML, then go over line by line and clean up the most possible HTML tags for section titles, and then extract all lines with the same HTML tags. Finally, we separated the whole HTML files by all the lines with the HTML tags of the section title.

3.5 Format Conversion (PDF)

Conversion to HTML and extraction method above are for resumes originally in docx format. However, for original pdf resumes, conversion to plain text was used.

3.6 Section Extraction (PDF)

To extract section information from plain text[12], we used string matching here, which means we write string matching rules and regular expressions to catch the key words that may be used as a title. Then, we separate the whole text by all the extracted titles.

3.7 Extraction Accuracy

By manually going over two hundred resumes to check how accurate the conversion and extraction are, the accuracy of extraction was up to 97%, so conversion to HTML and plain text was effective and the algorithm to extract information from HTML and plain text was efficient.

3.8 Preprocessed Dataset

After preprocessing the resumes, all resumes are in the form of an ID and their content information about each section. For each section, three pieces of information on the section name, the section content, and the lines that corresponding to the content are extracted. For section name, there are 'Education', 'Work Experience', 'Activities', 'Skills', 'Profile, and 'Other'. The section content is the text content under one section and the corresponding

lines are the content about all the lines in the section content and the tokenization form of the lines. In other words, the preprocess data is like this. See figure 3.1.

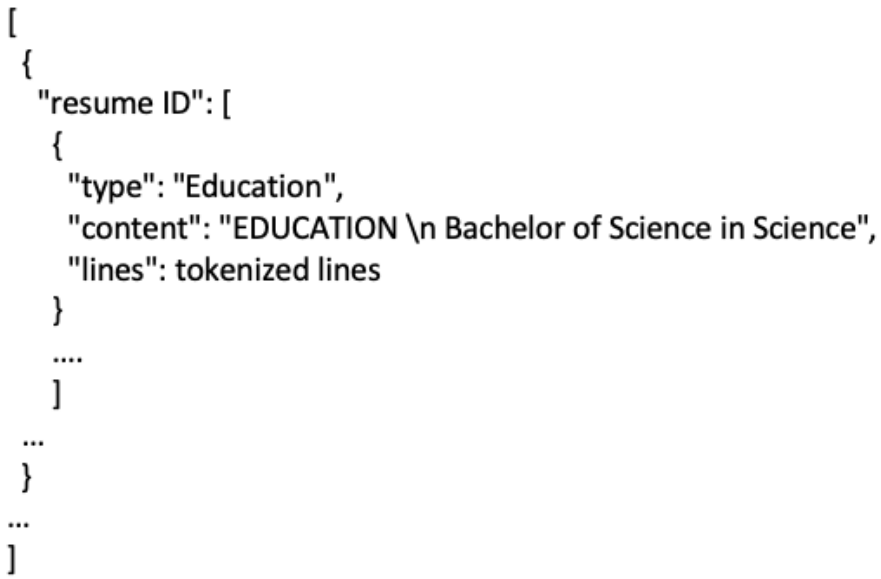

Figure 3.1: Cleaned Resume Content corresponding to different sections

Chapter 4

Approach

Approaches are various related to text classification. However, since the specialty of this dataset that every job has its own job admission standards, rule-based method can be used before doing unsupervised learning. Rule based method is mainly about using string matching to catch the important information about the job requirement description, then grabbing key information [15]from the preprocessed resumes, then make comparison, and finally judge whether the resume has all the required information in the job requirement description (Section 4.1). In addition to rule-based methods, unsupervised learning methods can be used on feature vector, which is a vector with several dimension to represent the information of the key words in different sections (Section 4.2). Instead of representing key information of the sections, augmenting the matrix allows us to use matrices to represent the whole text information in each section, which is known as Bags of Words

method (Section 4.3).

Finally, by mentioning the three approaches, ensemble models are also worthy trying and will introduce more in detail about the decision on choosing the components in the ensemble models in Section 4.4.

4.1 String Matching

From the annotation done by Dr. Elaine Fisher, Rebecca Thomas, Charlie Williams and Sabrina Sabir, there is a guideline when they decided whether admitting or rejecting a application. This guideline has been reviewed many times and it stated all the requirements for different Clinical Research Coordinator positions: CRC I, II, III, IV. Sample formats of the requirements for a CRC position are like the following:

(1.) High School Diploma, GED or Program Certificate (CNA, MA, Phlebotomy, Lab Tech) AND 1-year experience in a clinical setting/ or clinical role.

(2.) Technical Diploma (LPN, Medical Assistant)

(3.) Associate Degree or 2 years of college and AND 1 year experience in a clinical setting/ or clinical role.

(4.) Bachelor's Degree in a scientific or health related field

(5.) Bachelor's Degree in a non-scientific, or non-health related field and AND 1 year experience in a clinical setting/g or clinical role.

(6.) Bachelor's Degree in a scientific or health related field and Master's Degree in a non-scientific field.

(7.) Bachelor's AND Master's Degree in a non-scientific field or health related field AND 1-year experience in a clinical setting/clinical role.

(8.) Master's Degree in a scientific or health related field.

Therefore, there are multiple standards for one CRC level and any applicant who fulfills one of the standards can be admitted. In addition, since every level of CRC has standards, some applicants may fulfill more than one level CRC position. For such cases, the highest level of CRC position that they could be admitted would be marked because in the annotation, the four annotators would mark the highest position they could offer for each applicant's resume. In the clearly stated standards, standards on education and work experience are the only two fields that the four annotators care about. Within education, whether every education degree is science-related degree or not is also one of the consideration. Within work experience, work experience type is the one they focus on. Therefore, cleaning of the sections has firstly been done by using regular expression and string matching.

For education, there are 7 types of degree: MD, PHD, Master, Bachelor, Associate, Technical Diploma. In addition, for each degree, it is either science or health related major or non-science or non-health related major. We use regular expressions to catch the degree and the corresponding major type. Regular expression is a sequence of characters that define a pattern. When using regular expression, all text information with the same defined pattern would be extracted. Therefore, after using that, a 7-by-1 vector is created with entries representing MD, PHD, Master, Bachelor, Associate Technical Diploma. For example, a vector [[ ],[ ],[‘master’, 1],[‘bachelor’,0],[ ],[ ],[ ]] means the applicant has a science and health related master degree and a non-science and non-health related bachelor degree.

For work experience, there are 5 types of experience: not clinical related experience, experience in a clinical setting, clinical research related experience, clinical intern and lab experience. Based on the guideline, each experience category has a strict and distinct description, so we use similar approaches as the previous to extract and decide the work experience type based on the information from the work experience section. For every work experience, the time period of that experience is also extracted based on a strict regular expression. Finally, a 5-by-1 vector would be created with entries representing

the year duration for each type of experience. For example, a vector [3.333, 0, 1, 0, 0] means the applicant has 3.333 years of not clinical related work experience, and 1 year of clinical research related work experience.

So far, we extract the education information and work experience information. Then, by writing multiple if-statements, we can decide whether a applicant's resume fulfill a certain standard within all standards for one CRC level. Take the previous example, one sample standard is like the following: Bachelor's Degree in a non-scientific, or non-health related field and AND 1 year experience in a clinical setting or clinical role. Then, if a resume has an education vector as [[ ],[ ],[ ],['bachelor',0],[ ],[ ],[ ]] and a work experience vector as [0, 1, 0, 0, 0], then it is a qualified resume.

The real experiment with this method and its result are presented in Section 5.2.

4.2 Feature Vector

Based on the previous cleaned vectors for education and work experience, further cleaning should be done to create a feature vector for one specific resume. We decided to combine the two vectors as the feature vector.

For Work Experience vector, we keep the same as the previous with each

entry representing the corresponding year duration for each experience type. For Education vector, for each entry, if the corresponding degree for this entry is mentioned in the resume, then it is marked as 1. If no information, then it is marked as 0. If it is mentioned and also mentioned as science and health related degree, then it is marked as 2.

After combining the two vectors, each resume has a 12-by-1 vector with first five entries representing the work experience information and the rest 7 representing the education information.

By having all the feature vectors for resumes, we can do machine learning based on several models which have performed well in multiclass classification: Logistic Regression[5][9], Random Forest[2][16], Gradient Boosting[4], and Support Vector Machine[6][1][10]. In this thesis, more experiment details and results are listed in Section 5.2.

4.3 Bags of Words

In this thesis, the last model is using bags of words. Bags of words has performed well in multiclass classification[8]. Bags of words means that a text is represented as a bag of its words, or the set of its words, disregarding its word order present in the text. Therefore, in this thesis, bags of words

is only used on Education and Work experience sections. In the first step, all resumes are truncated to texts including only texts from education and work experience sections. Then, out of 2025 resumes, the TF-IDF score for any word is calculated and it should work well in the method of BOW[17]. TF-IDF score is defined as term frequency-inverse document frequency and has the following mathematical formula:

Frequency in this document* log(Total Number of documents/ document frequency)

Here, term frequency means how many times a word present in one resume, and term frequency means how many documents contain such word. By using this method, ever resume can be represented by two vectors which represent all the words in education and work experience section. Since different resumes have different texts, the 100 most frequently used words are selected to represent the entries in the vector for both education and work experience vector. In other words, for education vector, the first entry represents the most frequent words among all resumes, the second entry represents the second most frequent words, and so on. For experience vector, it had the same algorithm. Therefore, after creating two vectors for each

resume, we can replace the whole resume text information with these two vectors.

Then, we can do further models as Logistic Regression, Random Forest and Gradient Boosting in Section 4.3 and more experiment details and results are listed in Section 5.2.

4.4 Ensemble Models

For every possible trial in Section 4.1, 4.2, 4.3, further ensemble models are used. Ensemble models are using different sub-models as the machine learning method to learn different information, and combine the final results from every sub-model.

Ensemble Model 1: Select one best model on work experience, education and work experience and education, respectively(total 3 models), then combine their prediction result, and vote for the final result.

Ensemble Model 2: From all models performing work experience, education and work experience and education, select top 3 models for each(total 9 models), then combine three models prediction result for each part, and vote

for the final result.

Ensemble Model 3: Combine the 9 models in ensemble 3, aggregate the prediction result and vote for the final prediction result.

Chapter 5

Experiments

For the experiments, all data were shuffled and redistributed (Section 5.1). By using string matching, feature vectors, bags of words and ensemble models, we got some results (Section 5.2). Error Analysis would also be listed and discussed (Section 5.3).

5.1 Data Split

For the experiments, all application resumes in the dataset are randomly redistributed as training (75%),development(10%), and test(15%) as shown in Table 5.1.

Set	Number of Resumes
Training	1518
Development	202
Test	305

Table 5.1: Data split for our experiments.

5.2 Results

For all four models, in case that either work experience or education may play minor role when selecting applicants, three performances have been tested by looking at the information from work experience, education and combination of work experience and education, respectively.

String Matching

By using string matching (Section 4.1) to catch pieces of information about word experience and education from both the resume and the guidelines provided by Dr. Elaine Fisher, Rebecca Thomas, Charlie Williams and Sabrina Sabir, each applicant can be judged whether admitted or not to any level of CRC job position by just looking at whether their word experience and education fulfill the requirement in the guidelines. The accuracy of using this model to predict the level of CRC for a resume is in Table 5.2

Model	Accuracy
String Matching(Only on work experience)	42.9
String Matching(Only on education)	1.4
String Matching(Both work experience and education)	55.0

Table 5.2: Accuracy by using string matching to compare information from resumes and requirements in guidelines.

From the previous table, we know that when the four admission officers hire applicants, the guideline is not strictly followed and to some extent, they put more weight on the work experience of an applicant.

Feature Vector

By extracting key information of education and work experience from each resume, the work experience information is in the form of a 5-by-1 vector with each entry representing a type of experience and its value means the duration of years, and the education information is in the form of a 7-by-1 vector with each entry representing a degree and its three values 0, 1, 2 representing no degree, non-science and non-health related degree, and science and health related degree, respectively.

The dataset is now in the form of feature vectors and the models training on the dataset are Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine.

The accuracy of using these models on feature vectors to predict the level of CRC for a resume is in Table 5.3

Model	Only on work experience	Only on education	Both work experience and education
Logistic Regression + Feature Vectors	60.39	60.89	61.39
Random Forest Feature Vectors	60.89±0.29	**61.39**	60.89
Gradient Boosting Feature Vectors	**61.39**	60.40	60.89
Support Vector Machine Feature Vectors	58.41	60.89	**61.88**

Table 5.3: Accuracy by using classification models on feature vectors.

***Random Forest Model Development** In random forest model, the number of estimators and maximum depth of the random forest are the two parameters that are worth tuning [11]. For different number of estimators as shown in figure 5.1

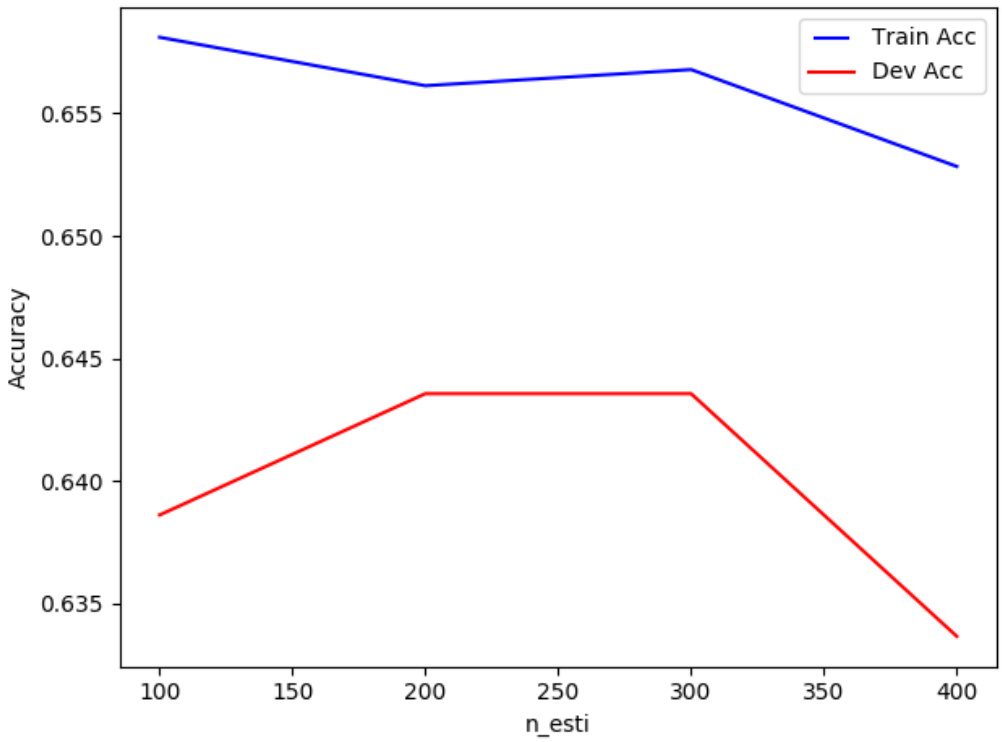

Figure 5.1: Accuracy vs Number of Estimators

development set accuracy is highest when number of estimators is around 200. When number of estimators is bigger than 200, the training accuracy is increasing while the development accuracy is decreasing, which means it probably is overfitting.

For different maximum depth of the random forest as shown in figure 5.2

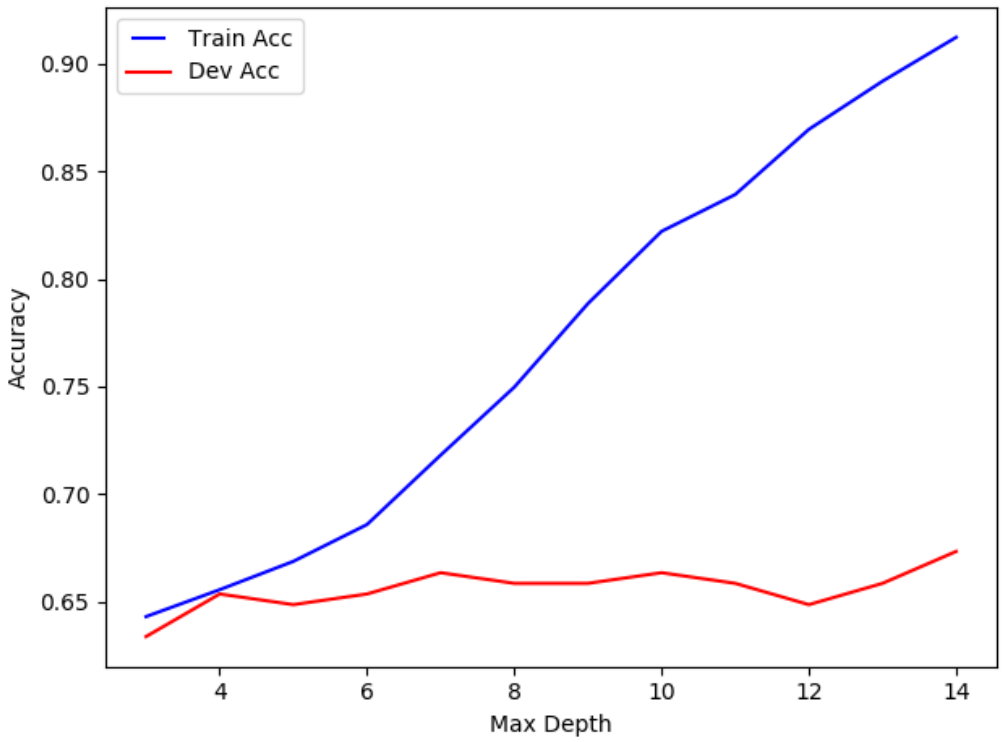

Figure 5.2: Accuracy vs Number of Maximum Depth

development set accuracy is highest when number of maximum depth is around 4. When it is bigger than 4, the training accuracy is increasing while the development accuracy is decreasing, which means it probably is overfitting.

***Gradient Boosting Model Development** In gradient boosting model[3], the number of estimators and learning rate are the two parameters that undergoes tuning in this thesis. For different number of estimators as shown in

figure 5.3

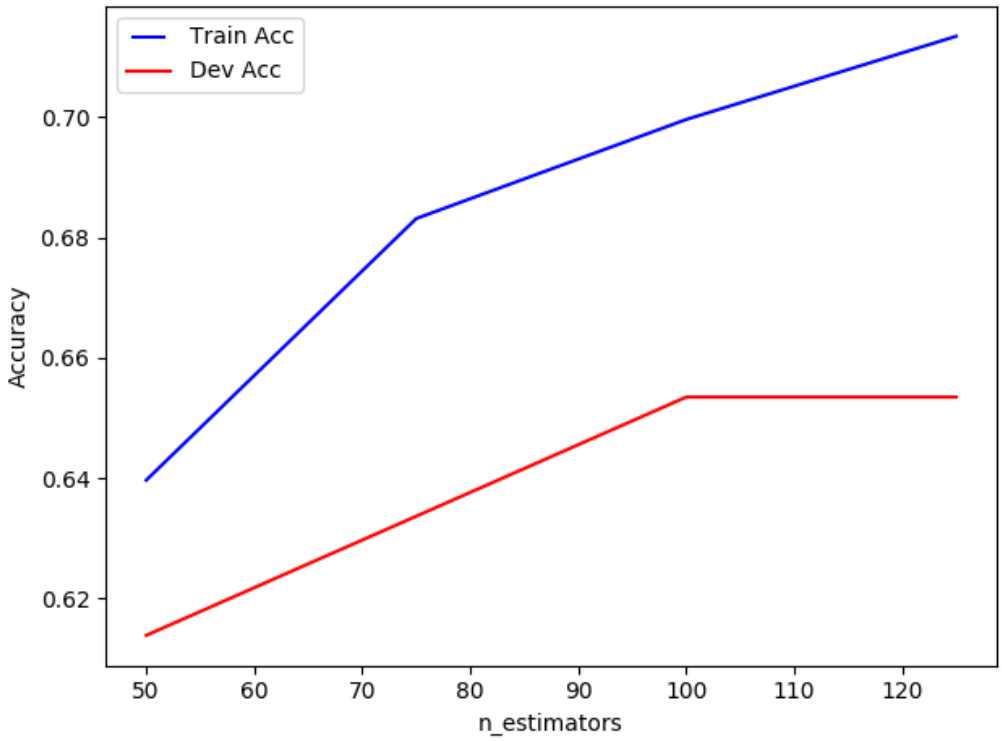

Figure 5.3: Accuracy vs Number of Estimators

development set accuracy is highest when number of estimators is around 100. When number of estimators is bigger than 100, the training accuracy is increasing while the development accuracy is decreasing, which means it probably is overfitting.

For different learning rate of the gradient boosting model as shown in figure 5.4

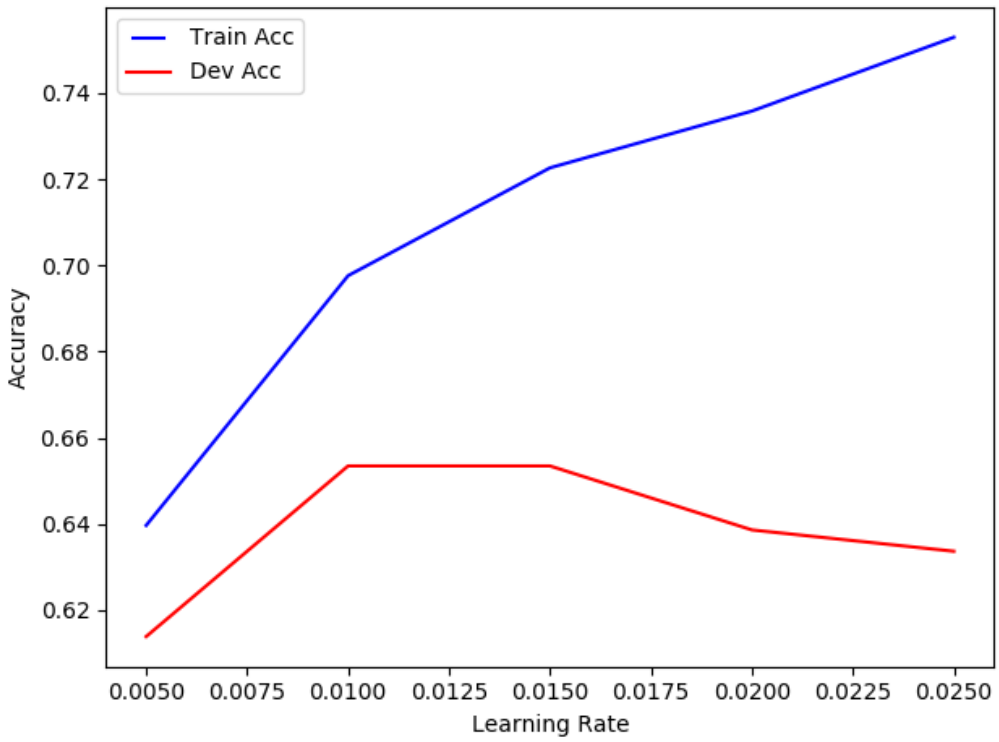

Figure 5.4: Accuracy vs Learning Rate

development set accuracy is highest when number of learning is around 0.01. When it is bigger than 0.01, the training accuracy is increasing while the development accuracy is decreasing, which means it probably is overfitting.

Bags of Words

Bags of words are using a 100-by-1 vector with each entry representing a word's TF-IDF score to represent the resume. The dataset is now in the form of feature vectors and the models training on the dataset are Logistic

Regression, Random Forest, Gradient Boosting, and Support Vector Machine.

The accuracy of using these models on BOW to predict the level of CRC for a resume is in Table 5.4

Model	Only on work experience	Only on education	Both work experience and education
Logistic Regression + Bags of Words	58.36	65.25	63.61
Random Forest + Bags of Words	**59.67**	65.25(±0.50)	63.61±0.19
Gradient Boosting Bags of Words	59.67(±0.19)	64.59	**65.34**
Support Vector Machine Bags of Words	58.69	**66.89**	61.97

Table 5.4: Accuracy by using classification models on Bags of Words.

Ensemble Models Ensemble modeling is a process combining multiple diverse models to predict an outcome, either by using many different modeling algorithms or using different training data sets.[14] The ensemble model aggregates the prediction from each model and results in once final prediction. Before ensembling models, the cleaned table of test accuracy is shown below in Figure 5.5

Model	Only on work experience	Only on education	Both work experience and education
Logistic Regression + Feature Vectors	60.39	60.89	61.39
Random Forest Feature Vectors	60.89±0.29	61.39	60.89
Gradient Boosting Feature Vectors	**61.39**	60.40	60.89
Support Vector Machine Feature Vectors	58.41	60.89	61.88
Logistic Regression + Bags of Words	58.36	65.25	63.61
Random Forest + Bags of Words	59.67	65.25(±0.50)	63.61(±0.19)
Gradient Boosting Bags of Words	59.67(±0.19)	64.59	**65.34**
Support Vector Machine Bags of Words	58.69	**66.89**	61.97

Table 5.5: Accuracy by using classification models on feature vectors and BOW.

For training models on different section: work experience, education, or work experience and education, the first ensemble model that we chose was training the models with best performance from each section, which is gradient boosting model on work experience using feature vector, support vector machine model on education using bags of words, and gradient boosting model on work experience and education using bags of words and then aggregate the results by voting.

See Table 5.6

Only on work experience	Only on education	Both work experience and education
Gradient Boosting+Feature Vector	Support Vector Machine+Bags of Words	Gradient Boosting+Bags of Words

Table 5.6: Model trained on different parts of resumes.

The next ensemble model was using top 3 models on each section, shown in Table 5.7

Only on work experience	Only on education	Both work experience and education
Logistic Regression+Feature Vector	Logistic Regression+Bags of Words	Logistic Regression+Bags of Words
Random Forest +Feature Vector	Random Forest+Bags of Words	Random Forest+Bags of Words
Gradient Boosting+Feature Vector	Support Vector Machine+Bags of Words	Gradient Boosting+Bags of Words

Table 5.7: Models trained on different parts of resumes.

The third ensemble model is to combining the previous nine models, and the result will be shown in the following.

With ensemble models, the **final result** is in Table 5.8 and Ensemble model 1 and Ensemble model 3 are listed in a separate Table 5.9

Model	Only on work experience	Only on education	Both work experience and education
Logistic Regression + Feature Vectors	60.39	60.89	61.39
Random Forest Feature Vectors	60.89±0.29	61.39	60.89
Gradient Boosting Feature Vectors	**61.39**	60.40	60.89
Support Vector Machine Feature Vectors	58.41	60.89	61.88
Logistic Regression + Bags of Words	58.36	65.25	63.61
Random Forest + Bags of Words	59.67	65.25(±0.50)	63.61(±0.19)
Gradient Boosting Bags of Words	59.67(±0.19)	64.59	**65.34**
Support Vector Machine Bags of Words	58.69	**66.89**	61.97
LG+Feature Vector RF+Feature Vector GB+Feature Vector	60.40		
LG+Bags of Words RF+Bags of Words SVM+Bags of Words		65.35	
LG+Bags of Words RF+Bags of Words GB+Bags of Words			61.39

Table 5.8: Accuracy by using classification models on feature vectors and BOW.

Model	Only on work experience	Only on education	Both work experience and education	Accuracy
Ensemble 1	Gradient Boosting+Feature Vector	Support Vector Machine+Bags of Words	Gradient Boosting+Bags of Words	61.39
Ensemble 2	Logistic Regression+Feature Vector Random Forest +Feature Vector Gradient Boosting+Feature Vector	Logistic Regression+Bags of Words Random Forest+Bags of Words Support Vector Machine+Bags of Words	Logistic Regression+Bags of Words Random Forest+Bags of Words Gradient Boosting+Bags of Words	61.88

Table 5.9: Ensemble Models trained on different parts of resumes.

5.3 Error Analysis

An error analysis is manually performed on 100 resumes. Errors mainly result from the following fields:

(1)Annotation Bias: 8 resumes may be annotated wrong based on the guideline.

(2)PDF File Problem: 1 scanned files in PDF format, whose content information cannot be extracted efficiently.

(3)DOCX File Problem: 2 files in DOCX format are using fancy structure that cannot be converted efficiently to HTML file. 1 file in DOCX format is using unorganized section titles: education title in picture format while work experience title in text format, which makes the HTML converter hard to detect titles.

(4)Writing Style Problem: 2 resumes do not follow common formats to write in education and work experience sections, which makes the information extraction process less efficient.

Given these error sources, some possible solutions are (1)Doing more rounds of annotation to ensure that the real classification strictly obey a certain guideline and rule.

(2)For file format issue, future researchers can come up with an approach to

use machine learning to learn the format of a resume so that section extraction can be more effective. (3)Using machine learning to learn the writing style may also be useful for section information extraction.

Chapter 6

Conclusion

This thesis introduces a newly-created dataset which includes real application resumes for Clinical Research Coordinator job position in School of Nursing at Emory University. This dataset is a complete dataset with original resumes from applicants and each resume has been clearly annotated. Since Clinical Research Coordinator is level-based job position with 4 different levels, the dataset contains 5 classes with 4-level classes and 1 not-admitted class. This dataset is different from the datasets in previous work, which mainly catergorizing resumes into different domains, but this thesis categorizes resumes under a same domain into different expertise levels. Designing a multi-class classification algorithm is performed by trying methods as string matching, feature vectors, bags of words and ensemble models. The best result reaches 66.89%, which is using support vector machine on education section.

Better result on this dataset is forecasting and promising in the future because there are many other models that can be used, like Bert, Albert and Roberta models in NLP field. In addition, for future work, more approaches can be taken to handle with the errors in this thesis. Giving more rounds of annotation can ensure a better annotation result and designing a new way to extract section information in the resume based on machine learning would make the extraction process more effective and less biased.

Appendix A

Complete Result

The **final result** including Ensemble model 2 is in Table A.1 and Ensemble model 1 and Ensemble model 3 are listed in a separate Table A.2

Model	Only on work experience	Only on education	Both work experience and education
Logistic Regression + Feature Vectors	60.39	60.89	61.39
Random Forest Feature Vectors	60.89±0.29	61.39	60.89
Gradient Boosting Feature Vectors	**61.39**	60.40	60.89
Support Vector Machine Feature Vectors	58.41	60.89	61.88
Logistic Regression + Bags of Words	58.36	65.25	63.61
Random Forest + Bags of Words	59.67	65.25(±0.50)	63.61±0.19
Gradient Boosting Bags of Words	59.67(±0.19)	64.59	**65.34**
Support Vector Machine Bags of Words	58.69	**66.89**	61.97
LG+Feature Vector RF+Feature Vector GB+Feature Vector	60.40		
LG+Bags of Words RF+Bags of Words SVM+Bags of Words		65.35	
LG+Bags of Words RF+Bags of Words GB+Bags of Words			61.39

Table A.1: Accuracy by using classification models on feature vectors and BOW.

Model	Only on work experience	Only on education	Both work experience and education	Accuracy
Ensemble 1	Gradient Boosting+Feature Vector	Support Vector Machine+Bags of Words	Gradient Boosting+Bags of Words	61.39
Ensemble 2	Logistic Regression+Feature Vector Random Forest +Feature Vector Gradient Boosting+Feature Vector	Logistic Regression+Bags of Words Random Forest+Bags of Words Support Vector Machine+Bags of Words	Logistic Regression+Bags of Words Random Forest+Bags of Words Gradient Boosting+Bags of Words	61.88

Table A.2: Ensemble Models trained on different parts of resumes.

Bibliography

[1] Yashima Ahuja and Sumit Kumar Yadav. Multiclass classification and support vector machine. *Global Journal of Computer Science and Technology Interdisciplinary*, 12(11):14–20, 2012.

[2] Ayo Akinyelu and Aderemi Adewumi. Classification of phishing email using random forest machine learning technique. *Journal of Applied Mathematics*, 2014, 04 2014. doi: 10.1155/2014/425731.

[3] Elizabeth A Freeman, Gretchen G Moisen, John W Coulston, and Barry T Wilson. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. *Canadian Journal of Forest Research*, 46(3):323–339, 2016.

[4] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[5] Gilles Gasso. Logistic regression. 2019.

[6] K. Gayathri and A. Marimuthu. Text document pre-processing with the knn for classification using the svm. In *2013 7th International Conference on Intelligent Systems and Control (ISCO)*, pages 453–457, 2013.

[7] Suhas Gopalakrishna and Vijayaraghavan Varadharajan. Automated tool for resume classification using sementic analysis. *International Journal of Artificial Intelligence Applications*, 10:11–23, 01 2019. doi: 10.5121/ijaia.2019.10102.

[8] Peng Jin, Yue Zhang, Xingyuan Chen, and Yunqing Xia. Bag-of-embeddings for text classification. In *IJCAI*, 2016.

[9] Peter Karsmakers, Kristiaan Pelckmans, and Johan AK Suykens. Multi-class kernel logistic regression: a fixed-size implementation. In *2007 International Joint Conference on Neural Networks*, pages 1756–1761. IEEE, 2007.

[10] Ajay Mathur and Giles M Foody. Multiclass and binary svm classification: Implications for training and classification users. *IEEE Geoscience and remote sensing letters*, 5(2):241–245, 2008.

[11] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3):e1301, 2019.

[12] Abhishek Sainani and PK Reddy. Extracting special information to improve the efficiency of resume selection process. 2011.

[13] Luiza Sayfullina, Eric Malmi, Yiping Liao, and Alex Jung. Domain adaptation for resume classification using convolutional neural networks. *CoRR*, abs/1707.05576, 2017. URL `http://arxiv.org/abs/1707.05576`.

[14] Grigorios Tsoumakas and Ioannis Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. In *European conference on machine learning*, pages 406–417. Springer, 2007.

[15] Mark D Wasson, James S Wiltshire Jr, Donald Loritz, Steve Xu, Shian-Jung Dick Chen, Valentina Templar, and Eleni Koutsomitopoulou. System and method for extracting information from text using text annotation and fact extraction, March 22 2011. US Patent 7,912,705.

[16] Baoxun Xu, Xiufeng Guo, Yunming Ye, and Jiefeng Cheng. An improved

random forest classifier for text categorization. *JCP*, 7(12):2913–2920, 2012.

[17] Zhang Yun-tao, Gong Ling, and Wang Yong-cheng. An improved tf-idf approach for text classification. *Journal of Zhejiang University-Science A*, 6(1):49–55, 2005.