

**RECRUIT BEST CANDIDATES WITH MACHINE
LEARNING**

Selvantharajah Thushanth

168272D

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

June 2018

RECRUIT BEST CANDIDATES WITH MACHINE LEARNING

Selvantharajah Thushanth

168272D

Thesis submitted in partial fulfillment of the requirements for the degree Master of
Science in Computer Science specializing in Data Science, Engineering and
Analytics

Department of Computer Science and Engineering

University of Moratuwa
Sri Lanka

June 2018

DECLARATION

I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature:

Date:.....

Selvantharajah Thushanth

The above candidate has carried out research for the Masters of Science thesis under my supervision.

Name of the supervisor: Dr. Charith Chitraranjan

Signature of the supervisor:

Date:

Abstract

In this research, I propose a robust approach for predicting personality traits of job candidates using machine learning. Relationship between personality traits and job performance has been studied extensively during the past few decades and thus this relationship can be utilized to overcome limitations in choosing the right candidates.

The proposed approach uses scenario-based analysis using machine learning techniques. Candidates will be asked to take part in scenario-based written conversations and their personality traits will be extracted from these conversations using machine learning techniques. Extracted personality traits of the candidates will be compared with the required job related characteristics in order to evaluate the fitness for the position for which candidates are applying. In order to categorize personality traits of candidates, the Five Factor model is used. Existing methods of evaluating personality traits such as standard set of questionnaires are susceptible to candidates providing false information and also time consuming.

Besides candidates' qualifications, knowledge and experience, candidates' personality traits also used to rank the candidates and shortlist them for face-to-face interviews. Thus, this technique not only allows recruiting right candidates to right position but also reduces significant amount of time and cost spent on evaluating candidates' suitability for given a job position by reducing the number of interviews to conduct. Further, this proposed system can be incorporated into existing e-recruitment system thus leveraging its effectiveness. Therefore, it is beneficial for companies since the proposed system helps to reduce cost and time consumption in the recruitment process while assisting them to choose more suitable candidates for a particular job position.

ACKNOWLEDGEMENTS

I would like to acknowledge the support and motivation given by Dr. Charith Chitraranjan, my supervisor, on successfully completing my research work. I also would like to thank him for his continuous suggestions and guidance throughout my research work. I wish to express my sincere gratitude to Dr. Surangika Ranathunga, for her valuable feedbacks during the initial phase of my research work.

Many thanks and appreciations go to all the staff members of Department of Computer Science and Engineering, University of Moratuwa for their generous help and motivation. Finally yet importantly, I would like to thank to my family for continuous support and encouragement.

TABLE OF CONTENTS

DECLARATION	i
ACKNOWLEDGEMENTS	iii
TABLE OF CONTENTS	iv
LIST OF FIGURES	vi
LIST OF TABLES	vii
LIST OF ABBREVIATIONS	viii
1. INTRODUCTION	1
1.1 Problem and motivation	1
1.2 Objectives	2
1.3 Organization of the thesis	2
2. LITERATURE REVIEW	4
2.1 Background	4
2.1.1 Recruitment process	4
2.1.2 E-recruitment systems	6
2.1.3 Personality models	7
2.1.4 Data mining	12
2.1.5 Supervised vs unsupervised learning	12
2.2 Traditional e-recruitment systems	13
2.3 Agent based e-recruitment systems	13
2.4 Personality based e-recruitment systems	15
2.5 Linguistic markers for personality traits	17
2.6 Predicting personality traits	19
2.7 Big five personality traits and job performance	26
2.8 Standard set of questionnaires	26

3. METHODOLOGY	28
3.1 Introduction	28
3.2 Focused personality traits	28
3.3 Data collection.....	28
3.3.1 Personality dataset	28
3.3.2 Yelp open dataset	30
3.4 Scenario based questions.....	32
3.5 Feature extraction	33
3.5.1 LIWC word category	33
3.5.2 Natural language processing technique	33
3.5.3 Process of assigning topics	38
3.6 Construction of the prediction model.....	42
3.6.1 Artificial neural networks	42
4. EXPERIMENTAL EVALUATION.....	45
4.1 Training and testing.....	45
4.2 Evaluation measures.....	47
4.3 Overall performance.....	49
4.3.1 Neural network model.....	49
4.3.2 Baseline method.....	50
4.3.3 Benchmark model	50
4.4 Overall comparison	51
4.5 Discussion	51
5. CONCLUSION AND FUTURE WORK.....	53
5.1 Conclusion.....	53
5.2 Future works.....	54
REFERENCES	55

LIST OF FIGURES

Figure 1.1: Traditional paper-based recruitment process using job advertising	5
Figure 1.2: The design and sequence of tasks in traditional paper-based recruitment process vs. the (new) recruitment process using e-recruitment	7
Figure 1.3: 16 personality types of the Myers-Briggs Type Indicator.	10
Figure 2.1: Profile for Personality Traits used in [27]	15
Figure 2.2: Top sixteen function words for low and high extraversion in stream-of-consciousness and deep self-analysis essays	23
Figure 2.3: Break down of accuracy rates achieved by each of the four feature sets for “Naïve Bayes” classifier [46]	25
Figure 3.1: Sample utterance from the data source	29
Figure 3.2: Personality scores for extraversion and naturalness by three individual human judges (namely userA, userB and userC) and average scores as well	29
Figure 3.3: Another sample utterance from the data source	30
Figure 3.4: Two independent human judges’ (namely userC and userD) personality scores and the average scores on each of the five personality traits in Big Five and naturalness	30
Figure 3.5: Shows Penn Treebank POS tags with examples (punctuation mark was also included)	36
Figure 3.6: Shows the structure of Artificial Neural Network	42
Figure 3.7: Shows all the main processes involved in measuring degree of positive emotions and negative emotions in a candidate’s answer	43
Figure 3.8: Shows how candidates’ personality evaluations can be incorporated into existing e-recruitment system	44
Figure 4.1: Bar chart shows the performance of the models using RMSE	51

LIST OF TABLES

Table 2.1: Shows the main eight categories used by [33]	21
Table 3-1: Shows four sample review texts extracted from Yelp Open Dataset with user given review scores	31
Table 3-2: Shows the word category used from LIWC utility	33
Table 3-3: Sample output from Stanford Log-linear POS Tagger	35
Table 3-4: Shows the seven content labels which are used to tag	39
Table 3-5: Shows how the text is tagged with content labels	40
Table 3-6: Describes five broad categories (or contents) a typical candidate's answer may contain	41
Table 4-1: Table shows the predictive input feature sets for extraversion and neuroticism	46
Table 4-2: Table shows the combination of hyper-parameter values used to construct artificial neural networks for extraversion and neuroticism	47
Table 4-3: Shows the results for extraversion and neuroticism from neural network model	49
Table 4-4: Shows the results for extraversion using 100 test set	49
Table 4-5: Evaluation results for neuroticism when using 10--fold cross validation	49
Table 4-6: Table shows the baseline results for extraversion and neuroticism	50
Table 4-7: Shows the result for extraversion and neuroticism using model from [39]	50

LIST OF ABBREVIATIONS

Abbreviation	Description
MAE	Mean Absolute Error
MBTI [®]	Myers-Briggs Type Indicator [®]
MSE	Mean Squared Error
POS	Part-Of-Speech
RAE	Relative Absolute Error
RMSE	Root Mean Squared Error
RSE	Relative Squared Error
SMO	Sequential Minimal Optimization