ENSEMBLE LEARNING BASED VOTING MODEL FOR DYNAMIC PROFILE CLASSIFICATION AND PROJECT ALLOTMENT

Suhas Tangadle Gopalakrishna and Vijayaraghavan Varadharajan

Infosys Limited, Bengaluru, India

ABSTRACT

Every year, lakhs of students right from college enter professional life through various recruitment activities conducted by the organization. The allotment of projects to the new recruits, carried out by the HR team of the organization is usually a manual affair. It is a time consuming and a tedious process as it involves manually opening each resume and analysing it one by one in order to assign a project. Companies round the globe are leveraging the power of artificial intelligence and machine learning to increase their productivity. In this paper, we present one such use case wherein artificial intelligence is leveraged by the organisation in allotment of projects to the new recruits. Current machine learning tools help in the allotment of projects to a few known popular domains on which the classifier has been trained explicitly. We tackle the problem with an ensemble learning based voting classifier consisting of 5 individual machine learning classifiers, voting to classify the profile of the candidate into the relevant domain. The knowledge extracted from the profiles for which there is no majority consensus among the individual classifiers is used to retrain the model. The proposed model achieves a higher accuracy in classifying resumes to proper domains than a standard machine learning classifier which is solely dependent on the training set for classification. Overall, emphasis is laid out on building a dynamic machine learning automation tool which is not solely dependent on the training data in allotment of projects to the new recruits.

KEYWORDS

Ensemble learning based voting classifier, Dynamic classification, Artificial Intelligence, Resume classifier, Association Rule Learning, Latent Dirichlet Allocation

1. INTRODUCTION

The jobs in the IT sector has been growing exponentially ever since it's inception. Companies round the globe recruit thousands of young talent every year. The new recruits have to be allotted projects by the organisation. Usually, HR team is entrusted with the responsibility of allotting the projects to fresh recruits, which is usually a manual affair. Manual allotment of projects to the new recruits by analysing the resumes of the candidates one by one is a tedious and a redundant process. In this paper, we have investigated a mechanism which helps the HR team in allotting the projects to the fresh recruits by considering the skill sets, interests and work experience mentioned in the resume of the candidates. The world of AI has grown significantly in the last decade or so. With the availability of large amounts of data, major tech companies round the globe are leveraging the concept of machine learning to increase productivity. According to recent survey by computer giant IBM [9], there is roughly about 2.5 Exabyte which corresponds

Natarajan Meghanathan et al. (Eds) : AIAA, WiMNeT, DPPR - 2018
pp. 77–88, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.81607

to about 2.5 billion gigabytes (GB) of data in the world. With the advent of big data tools like Hadoop, Spark, companies are capable of storing and analysing large amounts of data to build data hungry machine learning models for the purpose of automation. Enterprises these days are on a spree of expansion in new fields. Hence, they recruit individuals with varied specialization in order to gain traction in different fields. The current machine learning tools for resume classification do not entertain the resumes of candidates with specialization other than a few known, popular fields of domain on which it has been trained on. Thus, they fail to map the profiles with varied specialisation to projects. The main objective of our tool is to solve this issue by introducing the element of dynamism during classification of an instance of resume. This is achieved by building a model which is in a constant learning mode. For those resumes with specializations that has not been included in the training set, the tool extracts the features out of the resume, analyses and passes the knowledge extracted from the profile to a REST API trained on the stack-overflow dump for additional information on the category of domain to which the profile belongs to. Once the related topics based on the results from the Association Rule Mining of the Questions and Answers of the data dump is retrieved, it is analysed and subjected to topic modelling to extract relevant domain for the profile. The specialization of the resume is then added to the training set along with the extracted features. The process of training is then repeated in regular intervals in order to train the classifier on the new specializations. In this way, we eliminate the dependency of the model solely on the training set for categorization of resume and hence, eliminate the limitations of current machine learning models in the market specialized for this purpose.

2. Related Work

Considerable work has been accomplished in the field of text categorization by leveraging the concepts of machine learning in recent times [10]. In [1], the authors compare the different machine learning techniques which can be employed for text categorization [5]. But in this case, the categories into which the text would be classified is specified during the training phase. The model would then classify the text to a domain based on the training data alone. In [2] the authors proposea K-Nearest Neighbours machine learning classifier to classify a text into different clusters. V Ram and Prasanna have highlighted the importance of neural networks in the analysis of textual data [3]. They have argued that by feeding the neural networks with enough examples, the network would be able to predict the category of the text data in the test set more accurately than the conventional machine learning algorithms. Yieng Huang and Jingdeng Chen have argued for a deep learning model to classify text data [4]. Deep learning models which are a special variant of the neural networks discussed above are data hungry. These models have more than one hidden layers compared to conventional neural network models. Previously, research in the field of text categorisation has been solely dependent upon the training data. Our model aims at eliminating the sole dependency on the training set for mapping of profiles to the projects.

3. Our Approach

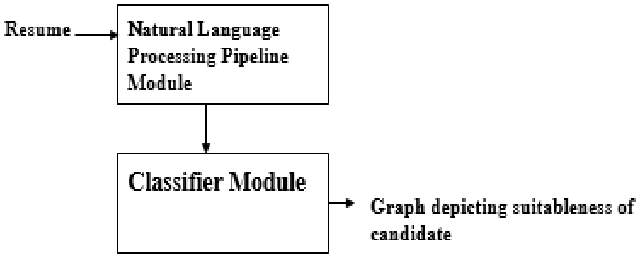

Resume
Natural Language Processing Pipeline Module
Classifier Module
Graph depicting suitableness of candidate

Figure 1. Block Diagram

The application developed comprises of two main modules as shown in Figure 1. The modules are: a) Natural Language Processing Pipeline (NLPP) and b) Classification module. The resumes of the new recruits are fed as an input to the NLPP by the HR team of the organisation. The NLPP module converts the resume into tokens that can be used by the classification module for the actual classification of the resume into a proper domain. The NLPP involves several steps which convert the resume input into tokens. The tokens generated are then fed as an input to the classification module. The classification module analyses the list of tokens generated by the NLPP module in order to allot a domain to the candidate. The application plots a graph depicting the relevance of the candidate with respect to various domains depending upon the knowledge extracted from the resume of the candidate. Hence, instead of only giving details about the most suitable domain for the resume, the relevance of the profile across various domains is plotted in the form of graph. The results from the graph aids the HR team of the organisation in allotting the projects to the candidates, eliminating the tedious process of opening each resume one by one and analysing them manually.

4. NATURAL LANGUAGE PROCESSING PIPELINE

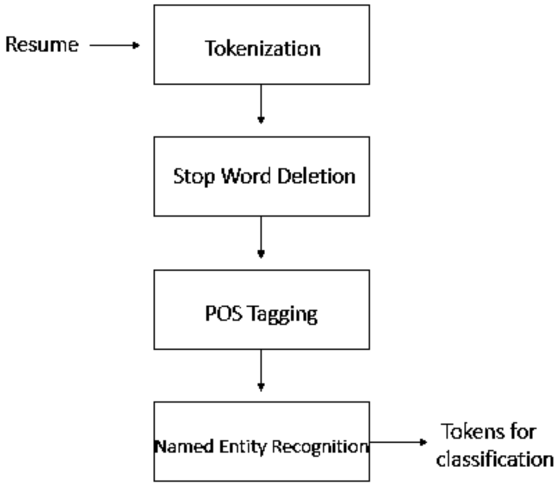
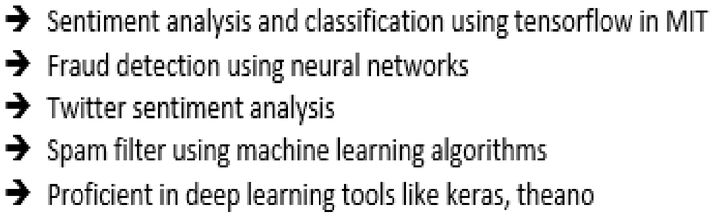


Figure 2: Flow chart of NLPP module

Figure 2 gives a detailed overview of the entire process of NLPP. The new recruits are allotted projects based on the content present in their respective resumes. The data in the resumes of the candidates are subjected to a NLPP in order to obtain only necessary and relevant details. Figure 3 depicts a portion of sample machine learning profile.



- ➔ Sentiment analysis and classification using tensorflow in MIT
- ➔ Fraud detection using neural networks
- ➔ Twitter sentiment analysis
- ➔ Spam filter using machine learning algorithms
- ➔ Proficient in deep learning tools like keras, theano

Figure 3: Portion of a sample machine learning profile

4.1. Tokenization

The sentences in the resume are segmented to obtain tokens. The delimiter for tokenization is the space character. Tokenization involves breaking up of the portion of the machine learning resume into tokens as depicted in Figure 4. The sentence in the resume, “sentiment analysis and classification using tensorflow in MIT”, is converted into a list of tokens comprising of individual words “sentiment”, “analysis”, “and”, “classification”,“using”,“tensorflow”,“MIT”.

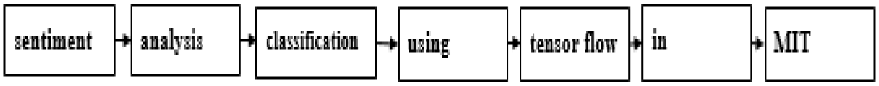


Figure 4: Tokens

4.2. Stop Word Deletion

Usually the resumes of the candidates are filled with redundant words such as 'is, and' etc. Such terms are called stop words. The removal of such simple stop words from the tokens obtained in the previous step resulted in a 28% rise in the efficiency of the classification model. The elimination is important because inclusion of stop words in the training set would result in false learning by the classifier, which would limit the efficiency of classification. The tokens, generated in the last step involve a stop word “and”. The word does not hold significance in the classification of the resume. Hence, the word is deleted from the tokens as shown in Figure 5.

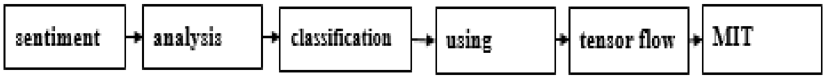


Figure 5: Tokens after stop word deletion

4.3. Parts of Speech Tagging

POS tagging is the next step followed in the NLPP. The step involves tagging the Part of Speeches to each of the tokens obtained after eliminating the stop words. English language has 8 different part of speeches: Verb, Noun, Adjective, Adverb, Pronoun, Preposition, Conjunction, Interjection. The tools and technologies used, Projects undertaken by the candidate is a noun or a pronoun. Since, the allotment of domain to the resume is dependent mainly on these features, only the tokens labelled as nouns and pronouns (NNP) are considered for the next step of NLPP. This step reduces the computation as the number of words considered for classification is reduced greatly. The POS tagging [7] of the tokens is shown in Figure 6. Only the tokens labelled as nouns move to the next stage in the pipeline.

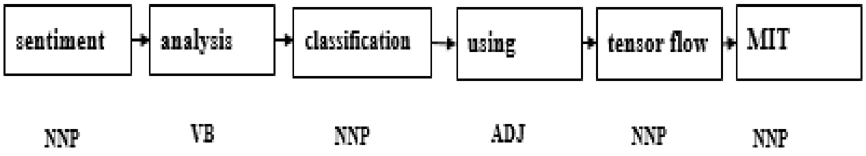


Figure 6: Tokens with POS tagging

NNP: Noun or a pronoun
ADJ: Adjective
VB: Verb

4.4. Named Entity Recognition

All the tokens labelled nouns and pronouns are subjected to Named Entity Recognition [8]. The tokens include the names of the candidate, educational institutions and place names. These tokens are eliminated in the present stage of the pipeline by identifying the candidate names, organizations and place name tokens, which are of trivial importance for allotment of a project to the candidate. This is the final step of the pre-processing pipeline. The tokens emerging out of this step are considered by the classifier for classification. The tokens which are recognized as name, place or organization in this step are eliminated as shown in figure 7. Only the contents shown in Figure 8 move to the classification module.

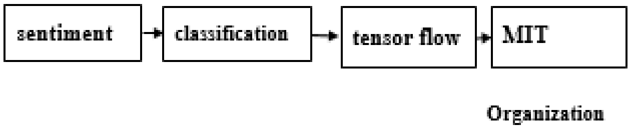


Figure 7: Tokens with Named Entity recognition

The process is repeated for every sentence in the resume portion shown in Figure 3.The final set of tokens are then forwarded to the classifier module as shown in Figure 8**.**

sentiment
classification
tensor flow

Figure 8: Output from NLPP

5. Classification Module

The classifier module receives the set of tokens from the NLPP module and is responsible for the classification of resumes into specific domains. The flowchart depicting the various steps involved in the module is shown in Figure 9. The classifier module is an ensemble learning [14] based voting classifier consisting of 5 different machine learning classifiers. The learning algorithms constituting the voting classifier are Naïve Bayes [6], Linear SVC, Bernoulli NB, Multinomial Naïve Bayes, Stochastic Gradient Descent classifiers respectively.

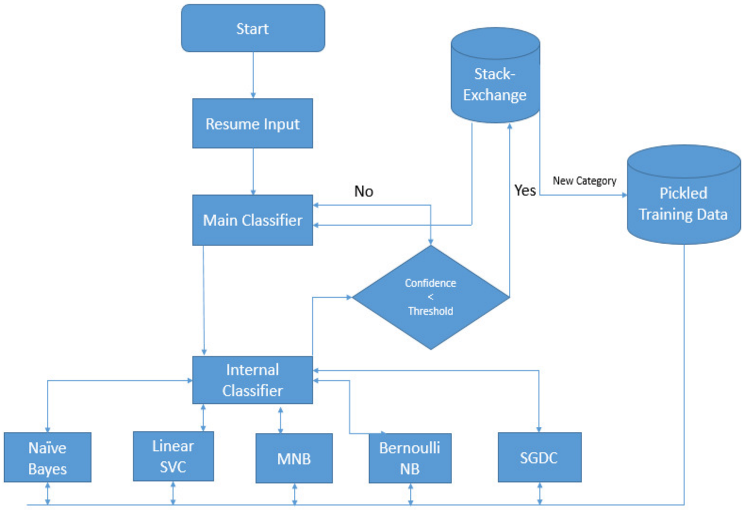


Figure 9: Flow chart of classification module

The set of individual learning classifiers constituting the ensemble learning based voting classifier are commonly used learning algorithms in the field of machine learning and form a benchmark for any new classifier. The initial set of domains identified are mentioned in Table 1. The training data set comprised of 30000 profiles of employee allocated to projects in the domains mentioned in table 1, based on their interests and work experiences previously. The data set formed the basis of the training set (supervised) of the individual machine learning classifiers constituting the ensemble learning based voting classifier. The dataset was divided into training set and a test set in 9:1 ratio. Hence, 27000 profiles formed the training set, while the rest 3000 formed the test set.

Table 1: Initial Domains to be mapped with the profile

Domains
Artificial Intelligence
Computer Architecture
Computer Graphics
Databases
Distributed Computing
Computer Networks
Web Technologies

Each classifier after training predicts the test data based on the learning from the training set. Efficiency of the classifier is the percentage of number of right predictions by the classifier on the test data. The efficiency of each individual classifier is shown in the Table 2.

Table 2: Efficiency of individual classifiers

Name of the Classifier	Efficiency of prediction in %
Naïve Bayes	79
Linear SVC	83
MNB	93
Bernoulli NB	89
Logistic Regression	81

The classifiers are allotted votes based on their respective efficiencies as shown in Table 2. This ensures higher influence of the classifier having greater efficiency in categorizing the tokens to a domain than the classifier having less efficiency, while mapping the profile to a domain. The number of votes given to each classifier is governed by equation 1.

$$\text{Number of votes to a classifier X} = \frac{Efficiency\ of\ classifier\ X}{Efficiency\ of\ classifier\ with\ least\ efficiency\ in\ training\ data} \qquad (1)$$

From Table 2, the least efficient classifier is Naïve Bayes (79% efficient), hence the denominator for the above equation 1, is 79%. Hence, for Linear SVC (83% efficient), the number of votes allotted is83%/79% = 1.05 votes. The votes allotted to individual classifiers is given in Table 3.

Table 3: Distribution of votes between different classifiers

Name of the Classifier	Number of Votes
Naïve Bayes	1
Linear SVC	1.05
MNB	1.17
Bernoulli NB	1.12
Logistic Regression	1.025

Confidence of the main classifier for the selection of domain for the set of tokens is the ratio of the number of votes casted in favour of the majority class to the total number of votes available with the individual classifiers.

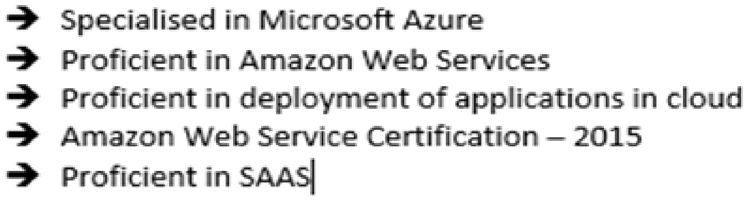
$$\text{Confidence} = \frac{\text{Number of votes cast in favor of majority domain}}{Total\ number\ of\ votes} \qquad (2)$$

Total number of votes cast by the classifiers are the sum of the votes allotted to the individual classifiers. From Table 3, the total votes allotted = 1 (Naïve Bayes) + 1.05 (Linear SVC) + 1.17(MNB) + 1.12(Bernoulli NB) + 1.025(Logistic Regression) = 5.365. If Naïve Bayes, Linear SVC and Bernouli NB choose machine learning domain for the set of tokens from Figure 8, while the rest of the classifiers chose distributed computing for the same set of tokens. Machine learning domain has a total vote share of 3.17, while distributed computing has a share of 2.195. Hence, the machine learning domain with a majority vote share of 3.17 is the domain selected by the main classifier. The confidence of the resume classifier from the equation 2 is given as 3.17/5.365 = 0.590.

Higher confidence of the model depicts higher consensus of the classifiers in selecting the domain and higher the chances of the classifier being correct in the decision to classify the set of tokens into the right domain. There are some instances, when the votes of the classifiers are cast equally on number of different categories. In this case, to break the deadlock, the domain which has been voted by the most efficient classifier is chosen.

6. Dynamic Classification

A threshold value is set on the confidence with which the resume is classified by the main ensemble learning based voting classifier. The threshold value set is 0.55, which is slightly greater than half of the total votes allotted to the classifiers. When the main classifier in the classification module predicts a set of tokens from the NLPP module with a confidence value less than the threshold value set, the classifier withholds the classification of the instance and calls the REST API trained on stack exchange dump to find a new category for the resume. A confidence value less than 0.55 signifies an absence of consensus among the individual classifiers in allotment of project to the resume.



➔ Specialised in Microsoft Azure
➔ Proficient in Amazon Web Services
➔ Proficient in deployment of applications in cloud
➔ Amazon Web Service Certification – 2015
➔ Proficient in SAAS

Figure 10: Portion of sample cloud computing resume

Figure 10 depicts a portion of sample cloud computing profile. Cloud computing does not come under any domain listed in Table 1. The sentence in the resume, "specialized in Microsoft Azure", forms a set of tokens after passing through the NLPP pipeline as shown in Figure 11.

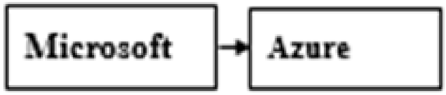


Figure 11) Tokens generated after NLPP

The individual classifiers listed in Table 2 are unable to classify the set of tokens generated by the NLPP pipeline because of the restricted knowledge of only those domains which are listed in Table 1. In such cases, the classifiers do not reach a consensus on the domain to be classified for the given resume. The model withholds the classification, calls the REST API which returns a plethora of related topics. The new topics returned from the API is based on the results from the Association Rule Learning carried out on the dump of Questions and Answers of Stack Overflow. Latent Dirichlet Allocation (LDA)is used for the topic modelling [12]. In natural language processing, LDA is a statistical model that allows sets of observations to be explained by unobserved groups that clarify why some parts of the data are similar.

Once the topic modelling is completed, the new domain is added to the list of domains in Table 1. The new data gathered in the process is added to the training set under the new domain. The internal models in Table 2 are then re-trained on the updated training set and the efficiency of the classifier is recalculated. A new domain of cloud computing is added to the Table 1, after topic modelling of the data returned from the REST API for the tokens shown in Figure 11.

7. RESULTS

Figure 12 depicts the output of the tool for the portion of machine learning resume shown in Figure 3.The model correctly classifies the instance of the resume into artificial intelligence domain. The classification is driven by the identification of deep learning tools like "tensorflow", "keras", "theano" and machine learning buzzwords like "spam filter", "analysis and classification", "fraud detection" in the resume.

Naïve Bayes, Linear SVC and Bernoulli NB classified the resume into AI domain while Multinomial NB and Logistic Regression classified the resume to Distributed computing and Computer Architecture respectively.

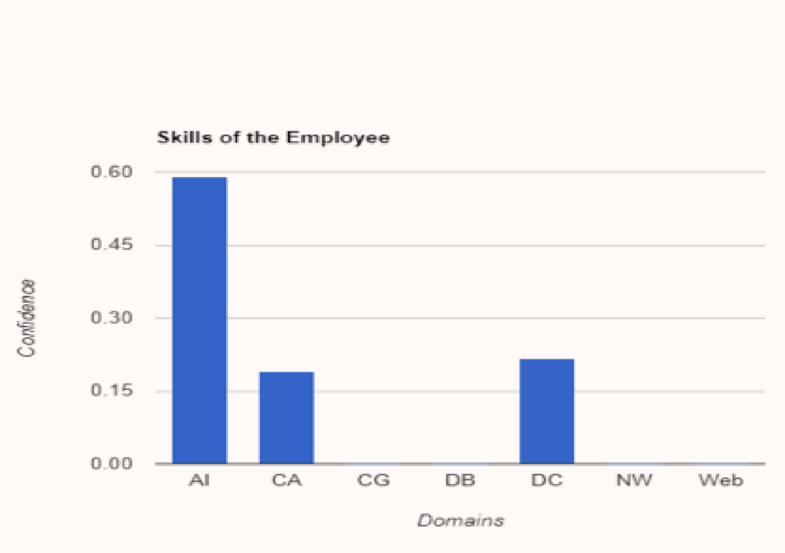


Figure 12: Output graph

AI: Artificial Intelligence
CA: Computer Architecture
CG: Computer Graphics
DB: Databases
DC: Distributed Computing
NW: Networks
Web: Web Technologies

.

The X axis forms the domains in the initial training set, while the Yaxis depicts the confidence of the main classifier across various domains. The profile is then allotted a project in the AI domain as classified by the classifier.

Confidence for AI = $\frac{\text{Number of votes for AI}}{Total\ number\ of\ votes}$

$= \frac{\text{Votes of Naive Bayes+Linear SVC+ BNB}}{Total\ number\ of\ votes}$

$= \frac{1+1.05+1.12}{5.365}$

= 0.590

Confidence for Distributed Computing = $\frac{\text{Number of votes for Distributed Computing}}{Total\ number\ of\ votes}$

$= \frac{\text{Votes of Multinomial NB}}{Total\ number\ of\ votes}$

$= \frac{1.17}{5.365}$

= 0.218

$$\text{Confidence for Computer Architecture} = \frac{\text{Number of votes for Computer Architecture}}{Total\ number\ of\ votes}$$

$$= \frac{\text{Votes of Logistic Regression}}{Total\ number\ of\ votes}$$

$$= \frac{1.025}{5.365}$$

$$= 0.191$$

The efficiency of the voting based classifier was 91.2%, predicting domains accurately for 2736 out of the total 3000 resumes in the test set, while in 80% of such cases, the confidence of the model was above 0.7. The accuracy without the re-training was 84.2%. Hence, the re-training after topic modelling of the related topics returned from the REST API increased the efficiency of the classifier by 8.3%.

$$\text{Increase in efficiency after retraining} = \frac{efficiency\ after\ retraining - efficiency\ before\ retraining}{efficiency\ before\ retraining} * 100 \quad (3)$$

$$= \frac{91.2-84.2}{84.2} * 100$$

$$= 8.3\%$$

8. CONCLUSION AND FUTURE WORK

The results from the model are encouraging. The promise of a tool which can help the HR team in making better decisions relating to the project allotment of a new recruit is met by the classification model. The ensemble learning based voting classifier performs extremely well compared to the individual classifiers while predicting most of the instances of the test data. This is because of the fact that the confidence of the model while categorizing resumes is influenced by the majority of the votes cast by the individual classifiers rather than a single classifier. An increase in the efficiency of classifier is observed on retraining with the information returned from the association rule mining [11] of stack-overflow questions and answers dataset. The future scope of the project is to build an ensembling deep neural network model [13] in place of the ensemble learning based classifier. Data hungry deep learning models especially Generative Adversarial Networks [15], can harness the huge amounts of unstructured and structured data available in the organisation, thus increasing the efficiency of the resume classification.

ACKNOWLEDGEMENTS

We would like to record our appreciation to all the people involved in the research and development of the resume classifier application. We express our gratitude to the In Step team at Infosys for providing us with the requisite dataset involving the resumes of the employees which formed the training set of the classifier.

REFERENCES

[1] A Comparative Study on Different Types of Approaches to Text Categorization, Pratiksha Y. Pawar and S. H. Gawande International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012

[2] Web Document Classification Based on Fuzzy k-NN Algorithm Juan Zhang Yi NiuHuabeiNie Computer and Information Computer and Information Computer and information China.

[3] V Ram, Prasanna, “A unique way of measuring the similarity of the documents using neural networks ,International Journal of Engineering Research and Development, Vol.2, no.6, July 2013, pp 397-401.

[4] Yiteng Huang, Jingdong Chen Blind, “Classifying the text using the power of deep learning”, International Journal of Engineering and Technology ,Taipei, Taiwan, 2013, pp 3153-3156.

[5] Daniel R and George V, “A Latent Semantic Analysis method to measure participation quality online forums”, 2016 IEEE 16th International Conference on Advanced Learning Technologies, January 2016, pp 108-113.

[6] Jongwoo Kim, Daniel X. Le, and George R. “NaïveBayesClassifier for Extracting Bibliographic Information from Biomedical Online Articles”, national Library of Medicine,8600 Rockville Pike, Bethesda, MD 20894, USA

[7] Natural Language Query Processing Using Semantic Grammar international Journal Of Computer Science And Engineering Vol II Issue II March 2010 pg no 219-233

[8] Natural Language Query Processing international Journal of Computer application And Engineering Technology and Science IJ-CA-ETS Oct 2009 pg no. 124-129

[9] https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/

[10] Karin Spark Jones, "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, vol. 28, no. 1, pp. 11-21, 1972.

[11] T. Griffiths, M. Steyvers, "Finding scientific topics", Proceedings of the National Academy of Sciences, vol. 101, pp. 5228-5235, 2004.

[12] D. M. Blei, A. Y. Ng, M. I. Jordan, "Latent Dirichlet allocation", J. Mach. Learn. Res., vol. 3, pp. 993-1022, 2003.

[13] Z.-H. Zhou, J. Wu, W. Tang, "Ensembling neural networks: Many could be better than all", Artificial Intelligence, vol. 137, no. 1-2, pp. 239-263, 2002.

[14] Z. Zhu, Q. Chen, Y. Zhao, "Ensemble dictionary learning for saliency detection Image And Vision Computing", Elsevier, vol. 32, pp. 180-188, 2014

[15] Lin Zhu, Yushi Chen, PedramGhamisi, JónBenediktsson, "Generative Adversarial Networks for Hyperspectral Image Classification", Geoscience Remote Sensing IEEE Transactions on, pp. 2018

AUTHORS

Suhas Tangadle Gopalakrishna:

Suhas works at Infosys Limited as a Specialist Programmer as part of Expert Track. His research interests include natural language processing, computer vision, human-computer interaction. He is active in various programming forums including Hacker Rank, Hacker Earth and Stack Overflow.

Dr Vijayaraghavan Varadharajan:

Vijayaraghavan is a Principal Research Scientist at Infosys Limited doing research in the field of data analytics, Searchable encryption, Security assessment, Cloud security, Authentication and Privacy protection. He has over 17+ years of experience in the fields of research, industry and academia. Prior to that he served as an Assistant Professor and guided many post graduate professional students. He has many granted US patents in key technology areas, published many research papers in International journals and conferences and also served as a Technical Reviewer, Program Committee member and Chair for many conferences around the globe.