Skill classified form job description

Anongphorn Janboonpeng 660632037,Putanyn Manee 660632067,Chattrapat Poonsin 660632076

Abstract

In today's dynamic job market, the process of matching job seekers with suitable employment opportunities has become increasingly complex.This paper was to analyze the classification of texts related to job.This study compiled job position qualifications from the LinkedIn job posting database.15887 samples were used to classify workgroups.Some typical parameters for job matching encompass the job title domain, job skills, and job description.Machine learning, with its remarkable capabilities in data analysis and pattern recognition, has emerged as a powerful tool to address this challenge.Logistic Regression and Naive Bayes(Multinomial) algorithm are used to classify resume into their respective categories.

Introduction and background

In today's competitive job market, the process of matching job seekers with suitable employment opportunities has become increasingly challenging.After overcoming the Covid crisis in 2022,affecting the labor market in Thailand is recovering quickly.People began to find new jobs and the company is looking for people to join the event.Our goal is to create a Job Matching System that simplifies the job search process. Instead of searching for jobs, job seekers will input their backgrounds, education, and skills.Simultaneously, employers can reach for candidates based on their specific needs,eliminating the need to wait for applicants to submit their resumes. Furthermore, in certain industries, employers may have the option to administer short exams, allowing them to identify job seekers with the precise skills they need.This project aims to contribute to this vital endeavor by leveraging cutting-edge technology and innovative strategies to create a more efficient and effective job matching System.

Literature review

From a research article by(Yasunobu Kino et al. ,2017 [1]) Job Matching is a means to connect a company and a candidate. There are two cases for job matching. Case 1 is to match a candidate to a specific job request, and case 2 is to match a job for a specific candidate.There is much numerical data and text data. Matching procedures are executed by many employees who hold the recruiter role using the search systems. The search systems generate a candidate long list at the request of the recruiter.

the research of Pradeep Kumar Roy et al.,[2] has classified the resume into different categories and recommends a resume based on the similarity index with the given job description with LinearSVM classifier. And had recommended that the performance of the model may be enhanced by utilizing the deep learning models like: Convolutional Neural Network,Recurrent Neural Network, or Long-Short Term Memory and

Others.

Abinash Tripathy et al.,[3] makes an attempt to classify movie reviews using various supervised machine learning algorithms, such as Naive Bayes (NB). by using the n-gram approach on the IMDb dataset.Found that as the value of ‘n’ in n-gram increases the classification accuracy decreases i.e., for unigram and bigram.

RESEARCH METHODOLOGY

The aim of this project is to reduce job searching time by Simplified search and quickly match the right candidate with employees so we must find algorithms for classification by using standardized CRISP-DM processes for managing data mining projects. It is divided into the following subcategories:

A.Business understanding

Data collection

The datasource : select Job posting and job skill of LinkedIn Job Postings Dataset from the kaggle website

Dataset Description: This dataset contains a nearly comprehensive record of 15,000+ job postings listed over the course of 2 days. Each individual posting contains 27 valuable columns and the job skill dataset contains 2 columns around 27,000+records.

The description and skill_abr are the key column for matching applicant qualifications.

B.Data Preprocessing

I. Text preprocessing description (job posting)

In this dataset ; description columns are applied to a bunch of text cleaning techniques to prepare the text for the next step, which is feature extraction.text preprocessing consist of clean contraction , remove junk characters as follows :

- Replace contraction mapping word
- Remove URL HTML and email ("<.*?>")
- Remove the special character ("!"#$%&'()*+, -./:;<=>?@[\]^_`{|}~")
- Remove New Line Code Snippet (\n)
- Remove Emoji (" ^^ ,- -,^3^")
- Remove Non-English Text
- Remove Digits
- Remove Punctuation
- **Stop words removal** : the general word which often shows in sentence but no meaning such as "a, an, the, of ".
- **Lemmatization** is a NLP technique used to reduce inflected words to their root word. It describes the algorithmic process of identifying an inflected word’s “lemma” (dictionary form) based on its intended meaning.
    - Transform original word to be the basic form, a word in the dictionary eliminate inflection of words

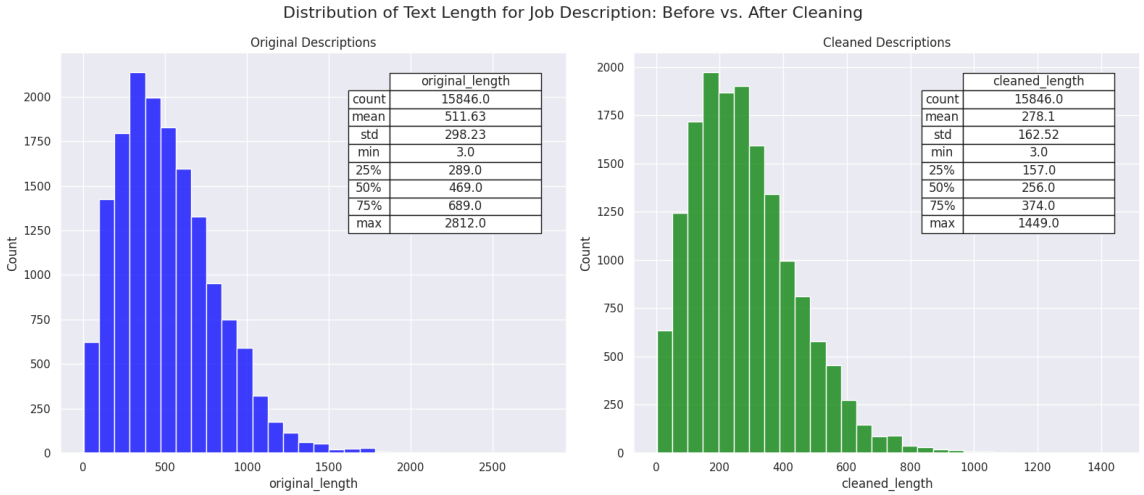


Fig.1 Before & after distribution of text length for job description

The visualization top word of job description after cleaning

Fig.2 Top word cloud visualization

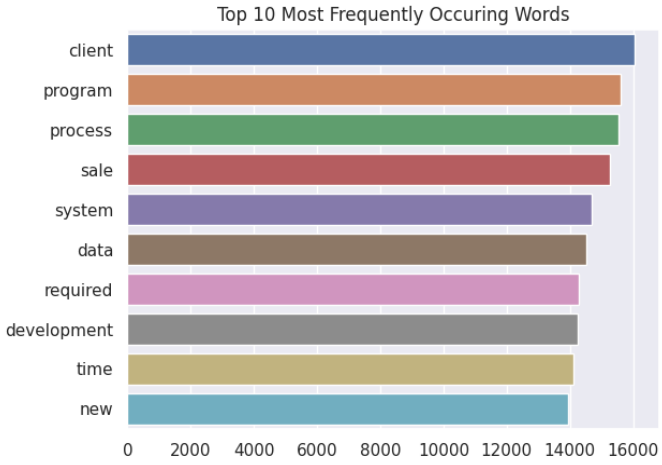


Fig.3 bar chart Top 10 most frequently word

II. Text preprocessing skill (job Skill)

35 job skill in dataset

Table 1. Job skill

1.ADM (Administration)	13.STRA (Strategy)	25.HCPR (Healthcare)
2.CNSL (Consulting)	14.SUPL (Supply Chain)	26.ADVR (Advertising)
3.HR (Human Resources)	15.BD (Business development)	27.MRKT (Marketing)
4.LGL (Legal)	16.GENB (General Business	28.PR (Public Relations)
5.PRJM (Project Manager)	17.ART (Art)	29.PRDM (Product Development)
6.MGMT (Management)	18.DSGN (Design)	30.ANLS (Analysis)
7.ACCT (Accounting)	19.WRT (Writing)	31.SCI (Science)
8.CUST (Customer Service)	20.EDU (Education)	32.RSCH (Research)
9.DIST (Distribution)	21.TRNG (Training)	33.QA (Quality Assurance)
10.FIN (Finance)	22.ENG (Engineering)	34.PROD (Production
11.PRCH (Purchasing)	23.IT (Information Technology)	35.OTHER (Other)
12.SALE (Sales)	24.MNFC (Manufacturing)	

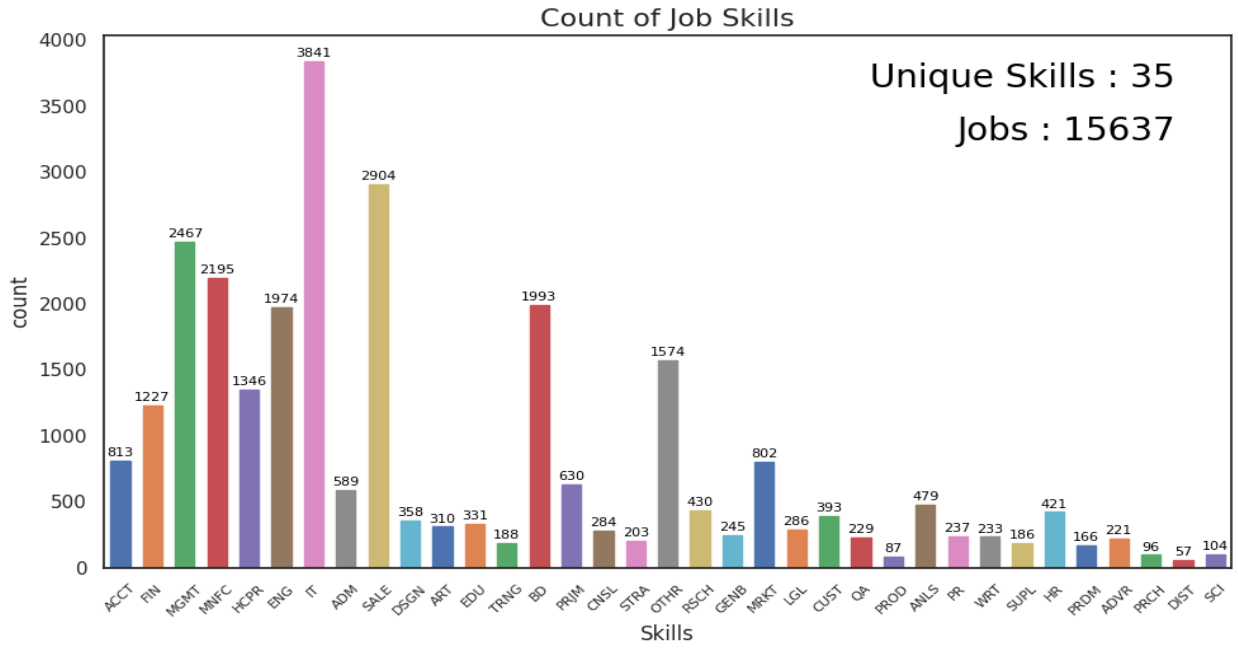


Fig.4 Count of job skills

The skills are grouped to be 10 skills 1 other by using the domain knowledge.

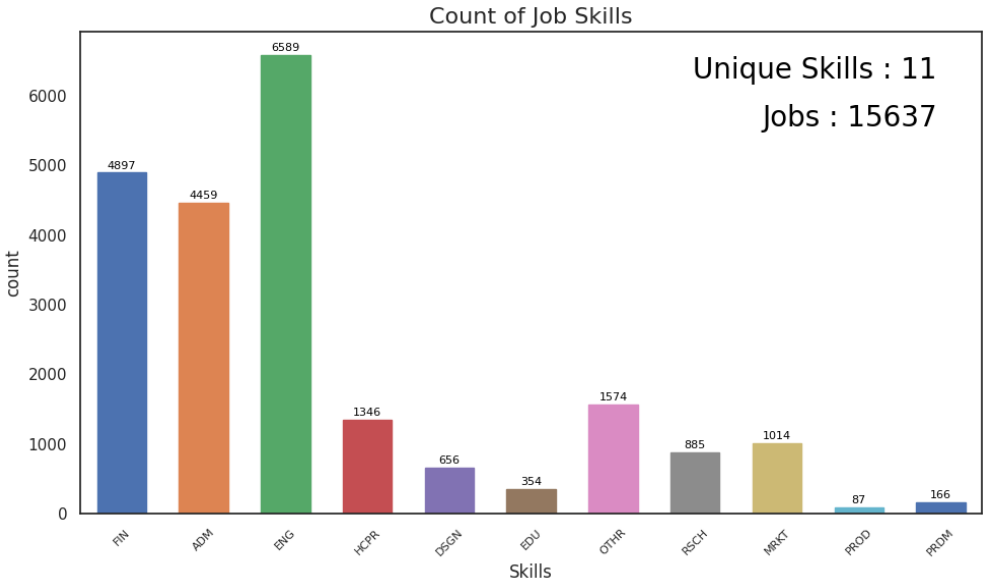


Fig.5 Count skill 11 group

Dataset are merged between Description & Skills ,shown as the fig.6
This fig.6 found 2 and 3 skill for in each job description but will select only one skills For classification.

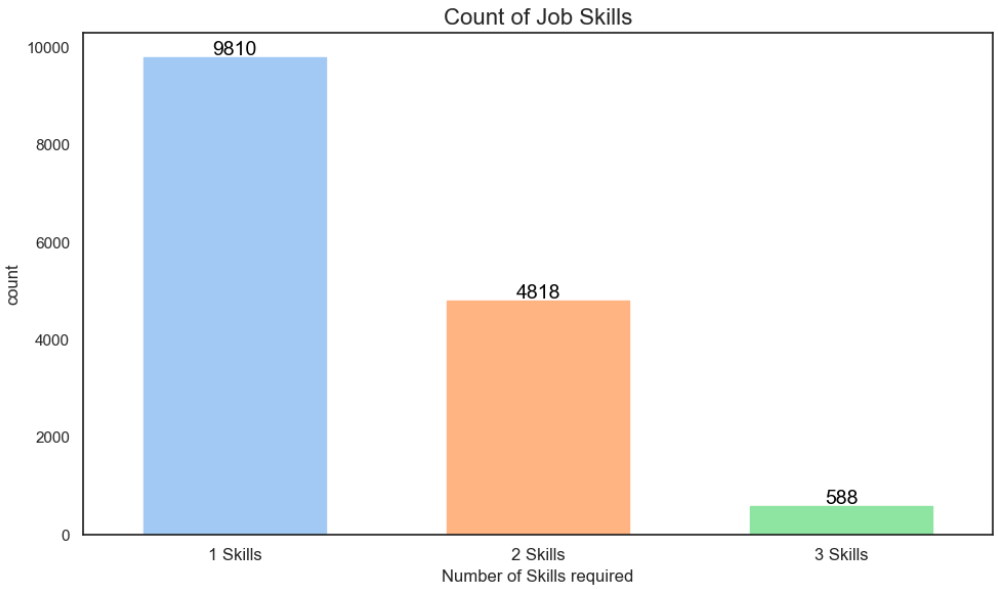


Fig.6 Number of skills required

The final data as the fig.5 ,next step dataset are applied models.

	job_id	title	description_cleaned	skill_abr_regroup	skill_count
0	85008768	Licensed Insurance Agent	many industry hurt last people still need insu...	FIN	1
1	133114754	Sales Manager	dynamic creative marketing professional lookin...	FIN	1
2	133196985	Model Risk Auditor	join u model risk auditor showcase financial a...	FIN	1
4	903408693	Office Associate	provide clerical administrative requestedanswe...	ADM	1
5	967848246	Education Manager	snapshot headquartered north miami paradigm se...	EDU	1
...	...	...	...	...	...
15212	3701373516	Sanitation Technician	locationwest columbia sc u type bakery operati...	ENG	1
15213	3701373522	Unit Secretary	title unit nursingreports nurse human date jun...	ADM	1
15214	3701373523	Radiology Aide, Perdiem	title radiology aide cat status nonexemptposit...	HCPR	1
15215	3701373524	MRI Manager	grade type officer date per week standard rang...	OTHR	1
15216	3701373527	Area Director of Business Development	health affiliate operate skilled nursing rehab...	FIN	1

9810 rows × 5 columns

Fig.7 Single-label dataset

C.Modeling

The subsequent stage involves researching various models for classifying job description.

The final dataset for modeling : Single skill data set

Step 1 : the dataset has multiple skill and description so transform by MultiLabelBinarizer

Step 2 : divided data into a training set (0.8) and a testing set (0.2) and random state 40 and applied ngam(1,2)

The ngram is referenced in the research of Abinash Tripathy et al.,[3] and fig.5 because mostly skills contain only 1 or 2 words so ngram (1,2) are enough.

Step 3 : Select model for classification in this project using Multinomial Naive Bayes and Logistic Regression Algorithm and parallel applying Bow and TF-IDF in each model.

Step 4 : show the performance model for classified text and matching job And comparison efficiency in each model.

Bag-of-Words(BoW) : Imagine you have a basket (or "bag") and every time you hear a word, you throw it into the basket. You don't care about the order of the words; you just care about which words are present and how many times each word appears. That's the essence of Bag-of-Words. It's a way to represent text by counting how often each word appears.

Example:

Sentence: "I love apples. I love oranges." BoW: {"I": 2, "love": 2, "apples": 1, "oranges": 1}

TF-IDF (Term Frequency-Inverse Document Frequency) : imagine you're in a library with many books. Some words, like "the" or "and", appear very often in many books, so they're not very special. But some words appear frequently in one book and rarely in others, making them more unique and important to that book. TF-IDF is a way to weigh words based on how frequently they appear in one document (like a book) compared to how frequently they appear in all documents (all books in the library). Words that are common in one book but rare in others get a higher weight.

Example:

In a library of 100 books, if the word "apple" appears 50 times in one book but rarely in other books, it will get a high TF-IDF score in that book.

In Simple Terms:

- BoW: Counts how often each word appears in a text.
- TF-IDF: Gives more importance to words that are frequent in one text but rare in others.

Multinomial Naive Bayes

Multinomial Naive Bayes is a specific variant of the Naive Bayes algorithm that is commonly used for text classification tasks.It's well-suited for problems where the features used for classification are discrete, such as word counts or term frequencies in text data. Multinomial Naive Bayes is widely applied in spam email detection, sentiment analysis, document categorization, and more.

The theorem formula is illustrated in Eq. (1). Letting that

A, B: events,
P: Probability,
P(A|B): Giving that B has occurred, what is the probability for event A,
P(B|A): Giving that A has occurred, what is the probability for event B,
and P(A), P(B): The independent probabilities of A and B.

$$p(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Logistic Regression Algorithm

Logistic regression algorithm is used to classify individuals in the categories based on logistic function. There are many instances when one does not get a perfect graph . For instance, we might encounter problems like the graph mentioned in Fig. 6a. The graph in Fig. 8a states how the action varies with respect to age. So, this graph is not at all appropriate and it does not fit all the data points. The solution is to implement a logistic regression algorithm. When we apply this algorithm to these data points, we get the graph as drawn in Fig. 8b. This graph is very much appropriate as it perfectly fits all the data points.[6]

Fig.8 a Graph when data points do not fit property. B Graph when logistic regression is applied and one gets a perfect curve.

Results and Discussions

The dataset (single label) is divided into training and testing and implementing classified models & Feature extraction.

Feature extraction

BoW model

The description is divided word by ngram and comparison word between 1-gram ,2-gram and 3-gram [fig.1]. n-grams of texts are extensively used in text mining and natural language processing tasks.when computing the n-grams you typically move one word forward for example ;

1-gram : ‘ I ’ , ‘ am ’ , ‘ a ’ , ‘ good ’ , ‘ girl ’ , ‘ friend ’

2-gram : ‘ I am ’ , ‘ a good ’ , ‘ girl friend ’

3-gram : ‘ I am a ’ , ‘ good girl friend ’

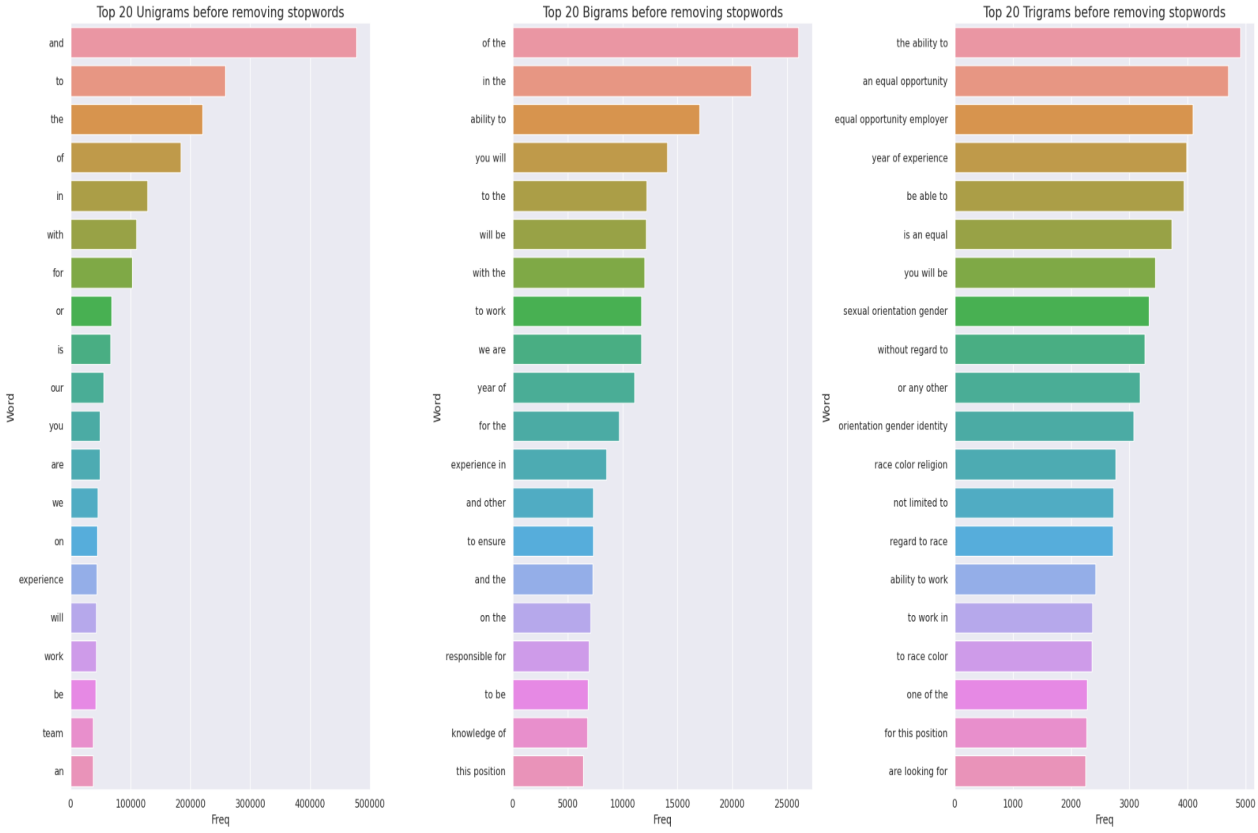


Fig.9 compare ranger of word after using 1-gram , 2-gram ,3-gram

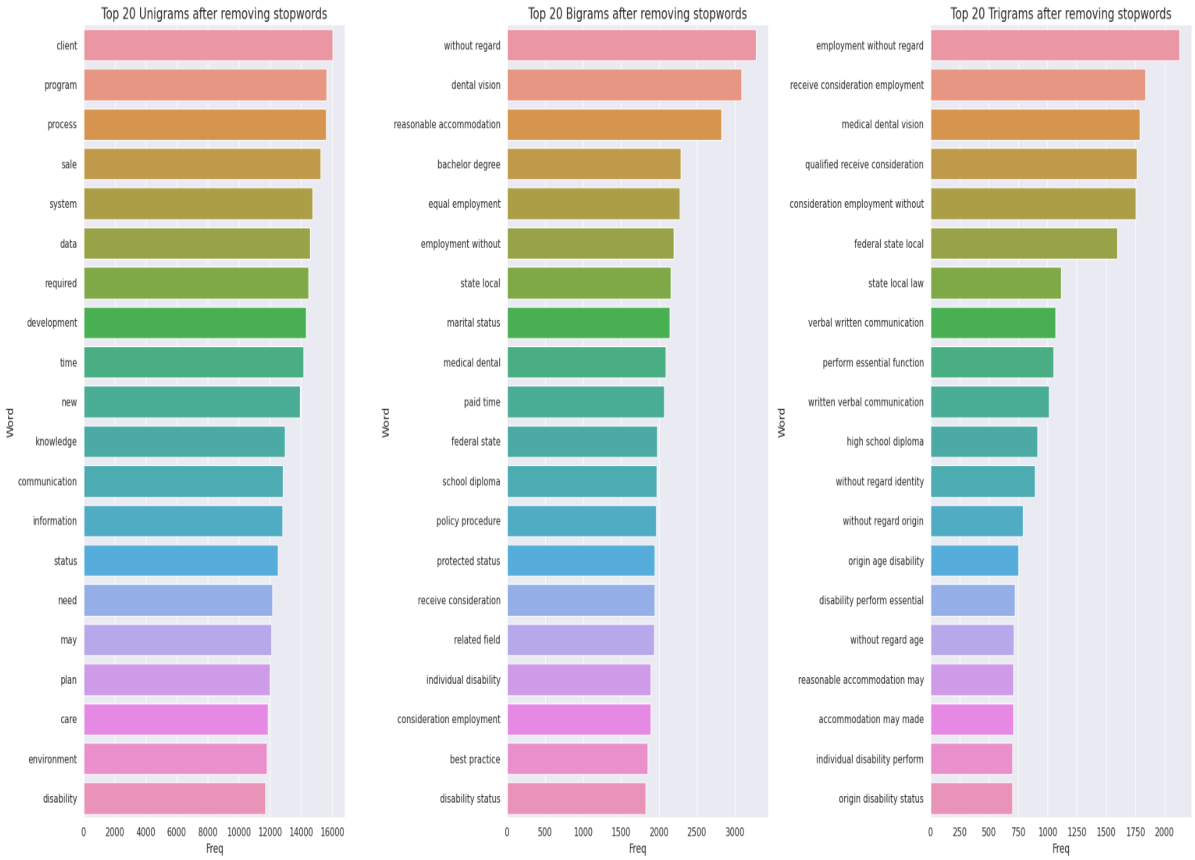


Fig.10 compare range of word which already removed stop words after using 1-gram , 2-gram ,3-gram

We can see that the system has trimmed down unnecessary words, leaving only important words.it can be seen that what people searching for employment want in a company will be related to customer,service,and business.

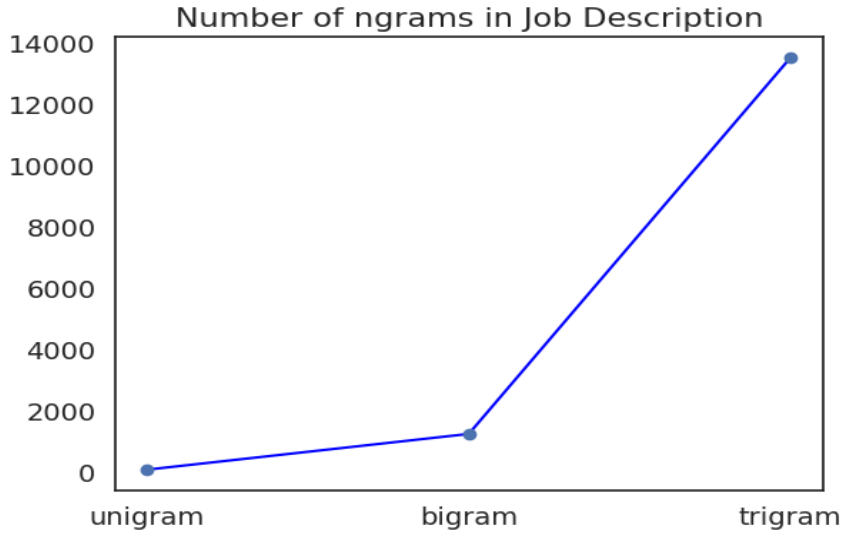
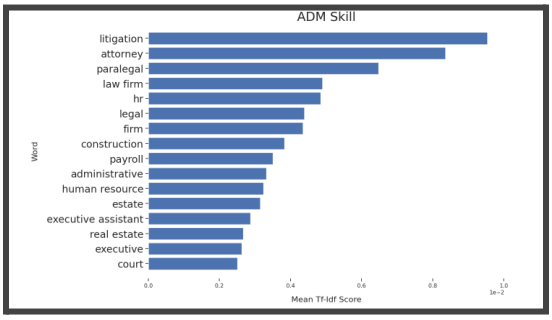
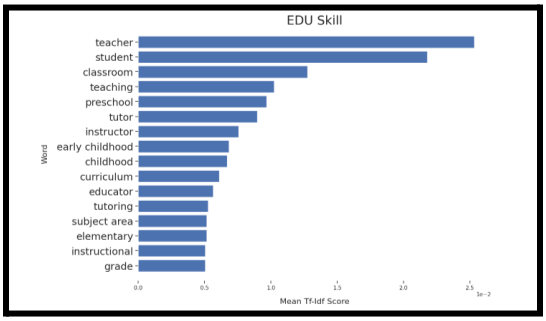


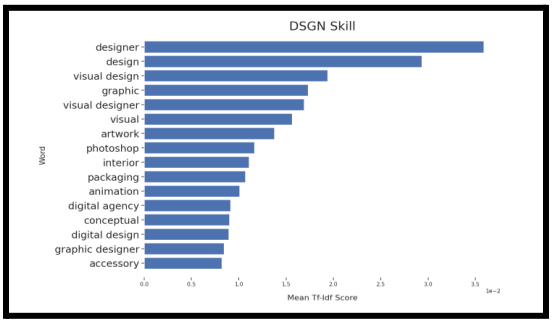
Fig.11 number of ngram in job description

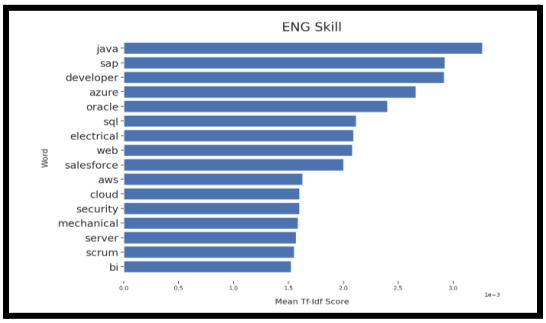
TF-IDF model

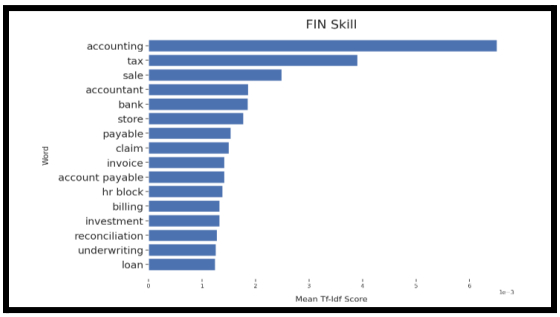
using TF-IDF techniques extract the main text using basic language processing and visualize TF-IDF Score of Job Description and in each Skills refer fig.9 found the top score TF-IDF is the same type of skill for example the EDU skill (education) including of description job ; teacher student classroom teaching etc.

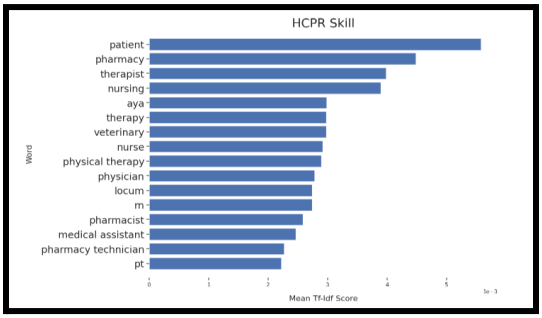


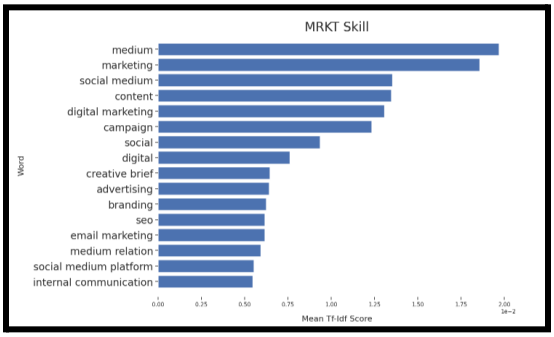


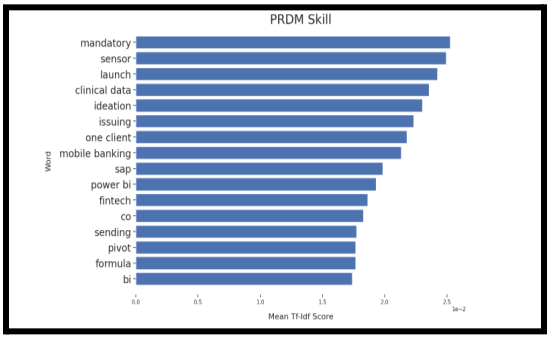












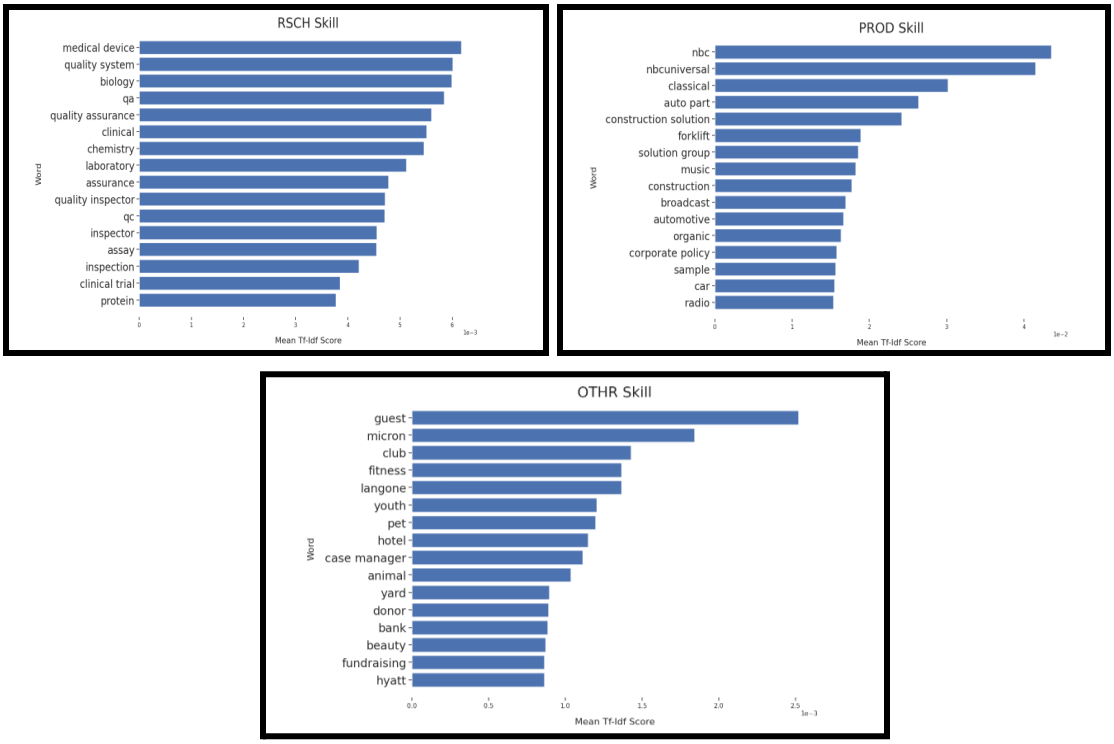


Fig.12 TF-IDF Score of Job Description and Skills

Single-label dataset
Logistic regression

TF-IDF technique accuracy 77.86%
BoW technique accuracy 81.64%

Table 2 comparison between TF-IDF and BoW technique of logistic regression model.

Skill	TF-IDF			BoW		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ADM	0.81	0.62	0.7	0.72	0.75	0.74
DSGN	1	0	0	0.67	0.33	0.44
EDU	1	0.29	0.45	0.75	0.71	0.73
ENG	0.8	0.91	0.85	0.84	0.88	0.86
FIN	0.78	0.94	0.85	0.88	0.9	0.89
HCPR	0.8	0.95	0.87	0.85	0.9	0.87
MRKT	1	0.095	0.17	0.64	0.33	0.44
OTHR	0.79	0.49	0.61	0.75	0.68	0.72
PRDM	1	0	0	1	0	0
PROD	1	0	0	1	0	0
RSCH	0.8	0.089	0.16	0.64	0.47	0.54

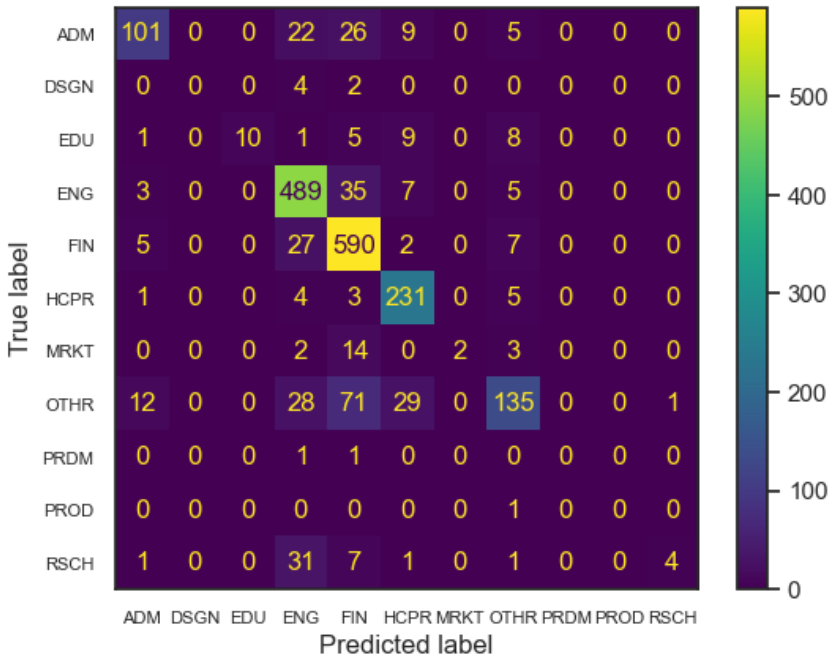


Fig.13 the confusion matrix of logistic regression with BoW

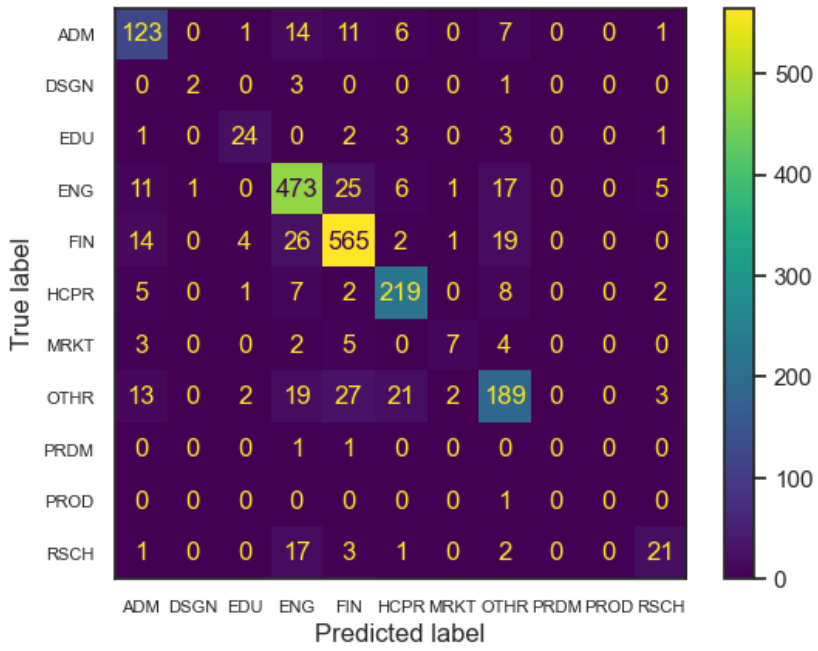


Fig.14 the confusion matrix of logistic regression with TF-IDF

Naive Bayes

TF-IDF technique accuracy 77.29%
BoW technique accuracy 72.50%

Table 3 comparison between TF-IDF and BoW technique of logistic regression model.

Skill	TF-IDF			BoW		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ADM	0.69	0.64	0.66	0.63	0.72	0.67
DSGN	1	0	0	1	0.17	0.29
EDU	1	0.21	0.34	0.67	0.71	0.69
ENG	0.78	0.9	0.84	0.84	0.88	0.86
FIN	0.79	0.9	0.84	0.86	0.87	0.86
HCPR	0.77	0.94	0.84	0.81	0.91	0.86
MRKT	1	0.095	0.17	0.88	0.33	0.48
OTHR	0.81	0.49	0.61	0.75	0.62	0.68
PRDM	1	0	0	1	0	0
PROD	1	0	0	1	0	0
RSCH	0.67	0.044	0.083	0.67	0.36	0.53

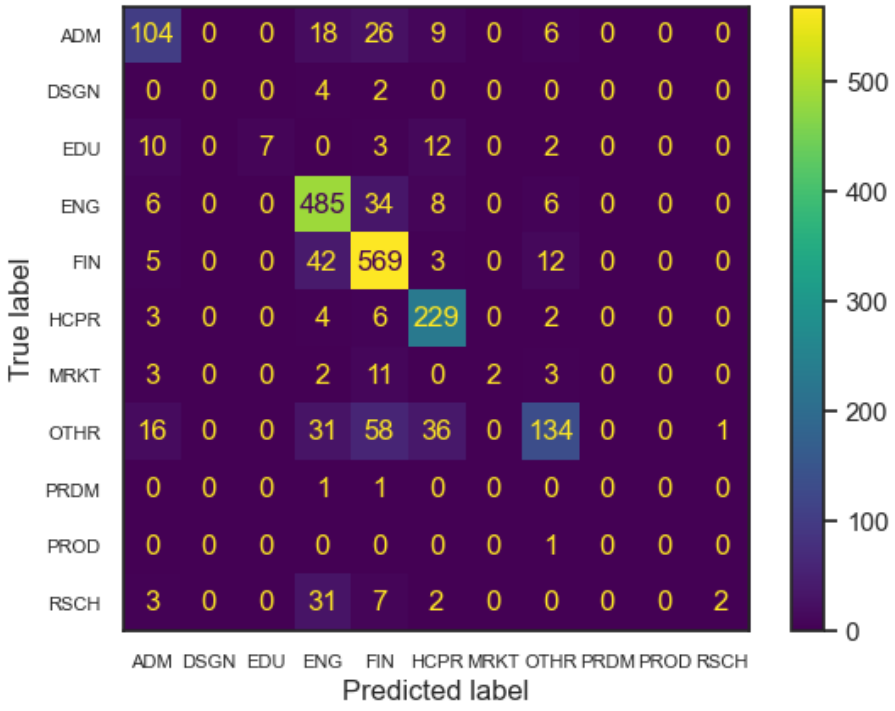


Fig.15 the confusion matrix of Naive bayes with Bow

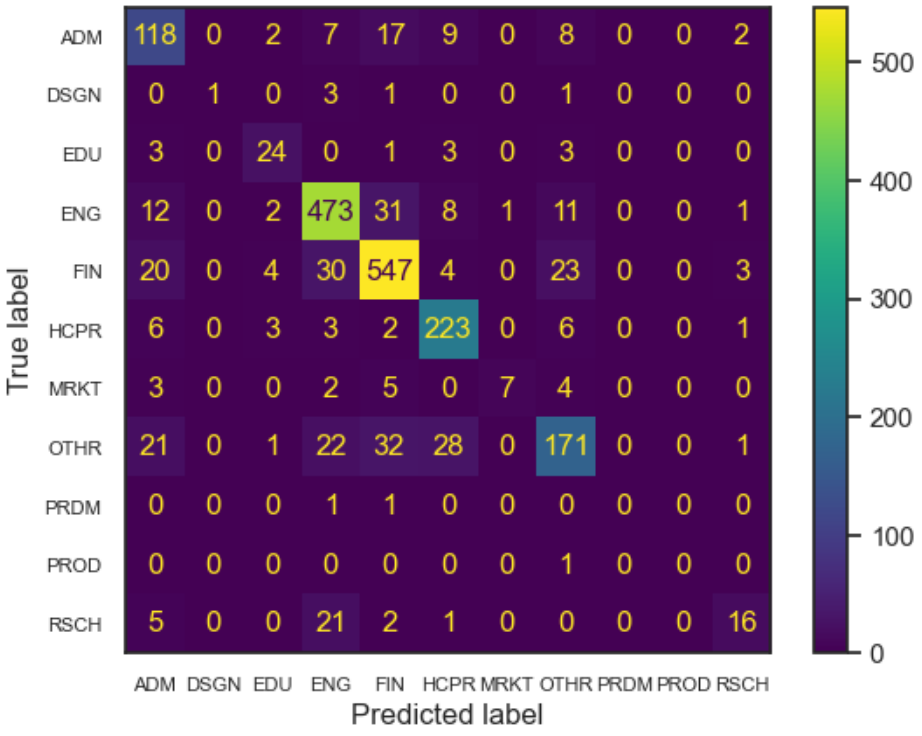


Fig.16 the confusion matrix of Naive bayes with TF-IDF

Single-label (remove insufficient skill label) dataset

Decided to remove insufficient skill that got a low f1-score because of lag of data from the previous model.

Logistic regression

TF-IDF technique accuracy 81.60%
BoW technique accuracy 83.95%

Table 4 comparison between TF-IDF and BoW technique of logistic regression model.

Skill	TF-IDF			BoW		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ADM	0.93	0.6	0.73	0.81	0.72	0.76
ENG	0.85	0.92	0.88	0.89	0.92	0.91
FIN	0.8	0.94	0.86	0.87	0.91	0.89
HCPR	0.83	0.92	0.88	0.87	0.89	0.88
OTHR	0.81	0.46	0.59	0.77	0.67	0.72

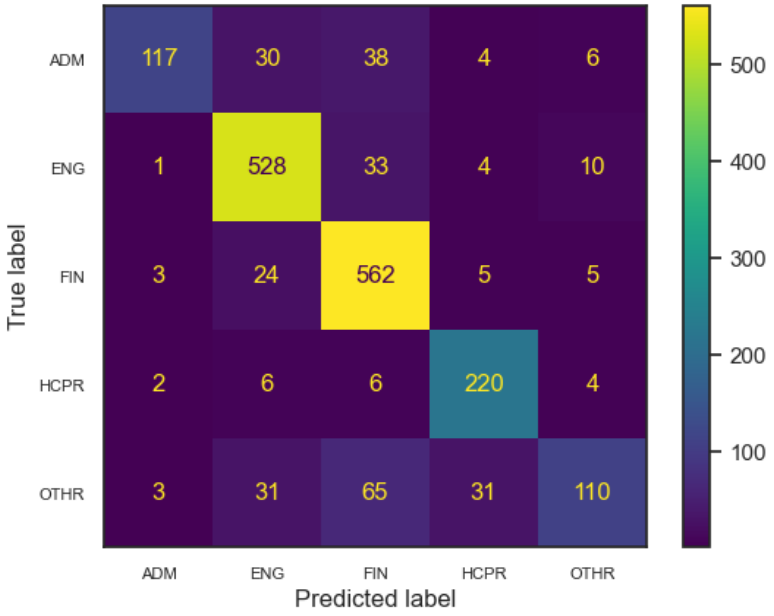


Fig.17 the confusion matrix of logistic regression with BoW

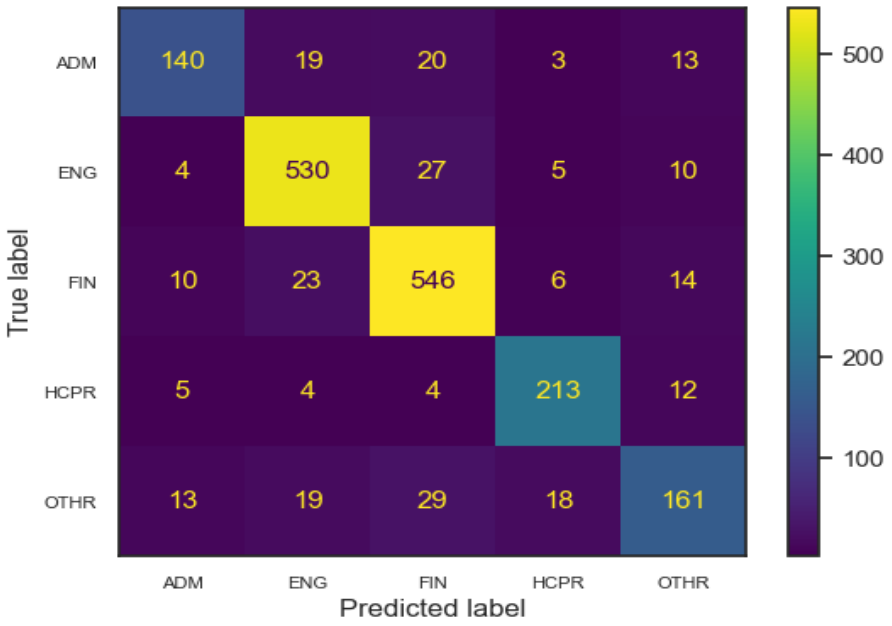


Fig.18 the confusion matrix of logistic regression with TF-IDF

Naive Bayes

TF-IDF technique accuracy 80.99%
BoW technique accuracy 75.44%

Table 5 comparison between TF-IDF and BoW technique of logistic regression model.

Skill	TF-IDF			BoW		
	Precision	Recall	F1-score	Precision	Recall	F1-score
ADM	0.83	0.64	0.72	0.72	0.72	0.72
ENG	0.84	0.91	0.88	0.88	0.9	0.89
FIN	0.8	0.9	0.85	0.87	0.89	0.88
HCPR	0.83	0.91	0.87	0.84	0.9	0.87
OTHR	0.81	0.47	0.6	0.77	0.61	0.68

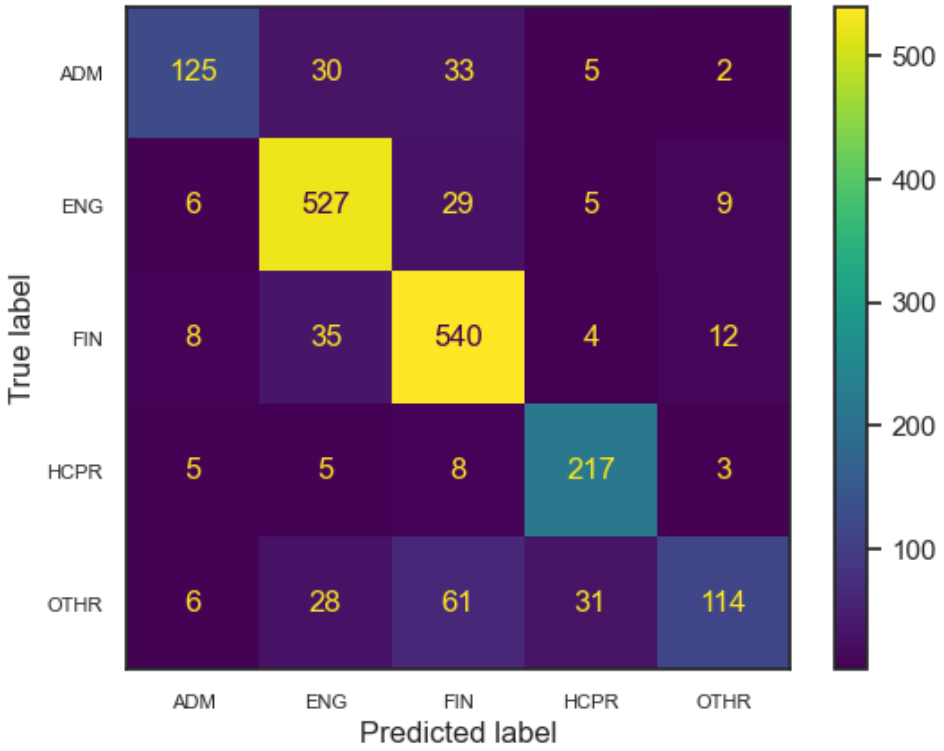


Fig.19 the confusion matrix of Naive bayes with Bow

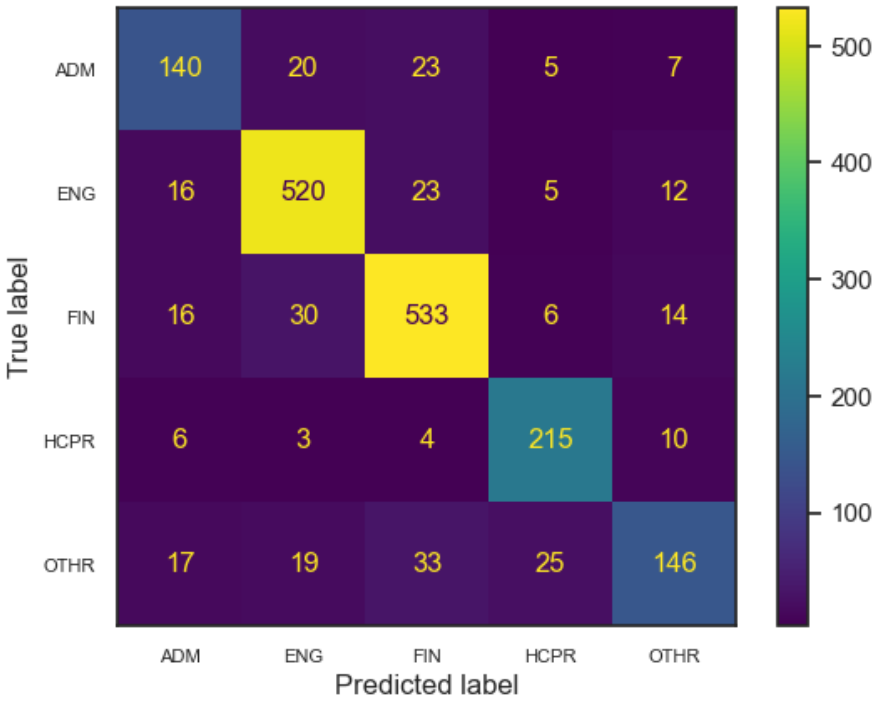


Fig.20 the confusion matrix of Naive bayes with TF-IDF

Model	TF - IDF	BoW
Logistic Regression	77.70	81.34
Naive Bayes	77.39	73.01

Fig.21 comparison accuracy single-label (11 Skill)

Model	TF - IDF	BoW
Logistic Regression	81.63	83.99
Naive Bayes	81.09	76.36

Fig.22 comparison accuracy single-label (5 Skill)

Conclusion

The project demonstrates the effectiveness of using Natural Language Processing techniques with logistic regression and naive bayes algorithms for job description classification . and recommends suitable candidates based on job description.after applied the models found that BoW is more efficient than TF-IDF after applying logistic regression but naive bayes which is extracted by TF-IF ,the accuracy are high efficiency.recommend utilizing removal of insufficiency skill label,The performance of the model may enhance and affect the extraction BoW with logistic regression have the highest accuracy 83.95%.

Future work

D.Evaluation

Evaluate the system's performance through extensive testing and Collect feedback from job seekers and employers for further improvements.

E.Deployment

Launch the Job Matching System as a user-friendly platform.Provide user support and continuously monitor the system's performance.

Reference

[1] Yasunobu Kinoa *, Hiroshi Kurokia , Tomomi Machidab , Norio Furuyab , Kanako Takanob. Text Analysis for Job Matching Quality Improvement.Procedia Computer Science 112 (2017) 1523–1530

[2] Pradeep Kumar Roya,∗ , Sarabjeet Singh Chowdharyb, Rocky Bhatia.A Machine Learning approach for automation of Resume Recommendation system in India.vellore Institute of Technology(2020) p.2324-2325.

[3] Abinash Tripathy∗, Ankit Agrawal, Santanu Kumar Rath.Classification of sentiment reviews using n-gram machine learning(2016) p.
approach

[4] RAJAT.(2023).Linkedin Job Position Dataset.Sep 3,2023,from https://www.kaggle.com/datasets/rajatraj0502/linkedin-job-2023 .

[5]Nasser, I., & Alzaanin, A.H. (2020). Machine learning and job posting classification: A comparative study. International Journal of Engineering and Information Systems (IJEAIS) ISSN, 6-14.

[6]Kanish Shah1 • Henil Patel1 • Devanshi Sanghvi1 • Manan Shah2(2020) Augmented Human Research (2020) 5:12 https://doi.org/10.1007/s41133-020-00032-0