Analysis of LinkedIn Jobs for Finding High Demand Job Trends Using Text Processing Techniques

[1]Abdul Karim Kazi, [2]Muhammad Umer Farooq, [3]Zainab Fatima, [4]Saman Hina, [5]Hasan Abid
karimkazi@neduet.edu.pk umer@neduet.edu.pk zainab.ned@cloud.neduet.edu.pk samhaque@neduet.edu.pk
hasanabid588@gmail.com
[1]Department of Computer Science and Information Technology, NED University, Karachi, Pakistan
[2]Department of Computer Science and Information Technology, NED University, Karachi, Pakistan
[3]Department of Software Engineering, NED University, Karachi, Pakistan
[4]Department of Computer Science and Information, Technology, NED University, Karachi, Pakistan
[5]Department of Computer Science and Information, Technology, NED University, Karachi, Pakistan

Abstract
LinkedIn is one of the most job hunting and career-growing applications in the world. There are a lot of opportunities and jobs available on LinkedIn. According to statistics, LinkedIn has 738M+ members. 14M+ open jobs on LinkedIn and 55M+ Companies listed on this mega-connected application. A lot of vacancies are available daily. LinkedIn data has been used for the research work carried out in this paper. This in turn can significantly tackle the challenges faced by LinkedIn and other job posting applications to improve the levels of jobs available in the industry. This research introduces Text Processing in natural language processing on datasets of LinkedIn which aims to find out the jobs that appear most in a month or/and year. Therefore, the large data became renewed into the required or needful source. This study thus uses Multinomial Naïve Bayes and Linear Support Vector Machine learning algorithms for text classification and developed a trained multilingual dataset. The results indicate the most needed job vacancies in any field. This will help students, job seekers, and entrepreneurs with their career decisions

Keywords:
Job posting; Natural Language Processing; Text Processing

I. Introduction

These job posting websites post thousands of jobs daily on their portal. But there is always a lack of such data through which we analyze what top jobs are posted on these websites daily, weekly, monthly, or in a year. So, text processing techniques should be applied there to process the jobs posted on that site. Text Processing includes Text classification. So here comes the text classification comes to the rescue. Using a text classifier, organizations can automatically structure those texts in a very fast and very less time-consuming way. With the help of text classification, the unstructured text can be categorized according to their categories, and then it will be helpful for the data analyzer to look into those data.

The text classification technique is used to categorize the text into organized groups. In this research paper, large data can automatically evaluate and then a proper pre-defined tag is assigned. Based on tags they generate the most relevant content of job descriptions according to their tags [1]. Then text preprocessing is applied to make data clean and useful for the machine learning algorithms. After preprocessing the cleaned data is provided to two machine learning methods for finding the high-demand jobs from LinkedIn. Multinomial Naïve Bayes and Linear Support Vector machine learning algorithms define the accuracy rate. Thus, this research work aims to classify the jobs data according to their categories also known as labels or tags. E.g., let's say there is a job description consisting of text related to Android jobs. So, the classifier will predict it as Android jobs. By using a machine learning algorithm, the trained model can predict.

LinkedIn is a great platform for finding or searching the jobs but when you want to find a job, you should have more than 2 to 3 skills. To address this problem, we did this research that defines the best way to find a job according to your skill. This research work provides a lot of benefits to students, job seekers, and entrepreneurs to find out which high-demand jobs are there in the industry. This will help the job seekers to decide what they can do to enhance their skills by knowing the prominent jobs.

According to the literature conducted in this research, several research works have been studied that were related to text preprocessing and text classification. It is a technique of tagging the data and then removing all duplicates and unnecessary data. This technique can be applied to small datasets as well as too big data sets. In the future, many researchers can work with this technique to solve a short problem [2].

In some research works, researchers work with big data which is also called the digital world. They collect a huge amount of data from different sites. They applied different machine learning algorithms to reveal a hidden pattern from the big data. Big data services application is used in businesses and to solve government, society, and science issues [3].

Manuscript received October 5, 2022
Manuscript revised October 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.10.29

Some researchers have also worked on text processing techniques to suggest top job postings from the website to help the user. Many researchers use this technique to automatically classify unstructured data and produce multilingual datasets. Some researchers use text processing techniques for finding the best employees for their organization. IJRM uses this technique for unstructured data to structured data and also multilingual datasets. They used the Support Vector Machine algorithm (SVM) and Naïve Bayes (NB) to acquire the best results. However, this research is for media platforms [4]. Some researchers conducted the research which can find the best employees for an organization. When the list of employees is given to the HR department, they can easily find out the best employee for their company [5].

In another research, the authors presented a classification system using descriptions (represented by their brief resumes) of LinkedIn profiles. These researchers have built a classifier that utilized 'Term Frequency Inverse Document Frequency- TFIDF' and 'Convolution Neural Network-CNN' for training the model. Using resumes of the applicants, they identified and classified the relevant information into related categories. In addition, these researchers also employed various algorithms with multiple combinations of features on the dataset of resumes. Support Vector Machine (SVM), Naïve Bayes (NB), and TFIDF were used as state-of-the-art methods. After that, the researchers used the above algorithms with another algorithm i.e., CNN which performed better as compared to the previous approaches.

Another group of researchers also investigated skill requirements for job positions related to data analysts and business analysts [6]. This paper also focused on job advertisement content but in a different manner. These researchers described a content analysis of jobs posted on an online website related to the above-mentioned positions. They present a rank-wise list containing skills categories for the studied position. They also work on another thing; they provide a pairwise comparison between data analyst and data scientist also on business analyst and business intelligence analyst. They have employed a support vector machine (SVM) and Multinomial Naïve Bayes for the classification of technical skills of the mentioned job categories.

This research paper aims to identify how well the two algorithms i.e., Support Vector Machine algorithm (SVM) and Naïve Bayes (NB) perform when providing the text (i.e., job description). This paper is the beginning of the research to apply machine learning algorithms for the categorization of multilingual job data. The obtained result can be used as a statistical analysis of job data.

As per the estimation of multiple data analyst, around 80% of the information related to high-demand jobs are in the form of unstructured data. The majority of this huge amount of data is in a textual form which is time-consuming and difficult to analyze and organize. Several types of research have been reported in a survey to analyze these big datasets in critical sectors such as financial sectors [7, 8]. In this survey, different applications that employed data mining techniques have been explored that use real-time big datasets from social media. As the amount of data on the websites is increasing every millisecond, trend analysis has gained importance in other fields as well. Researchers have analyzed big data related to the marketing domain using semi-automatic approaches of text mining [9]. They have used literature from 2010 to 2015 to analyze five dimensions in the marketing domain. Reported research concluded that there is still more contribution from the research community required to improve big data applications for this domain.

In the same context, online job advertisements were also analyzed through a semi-automatic approach. For this purpose, semantic structures of the dataset were studied using "latent Dirichlet allocation (LDA)" and probabilistic topic-modeling technique. This research contributed a systematic approach for industries to review updated job skills and improve competency levels while the recruitment process. LDA was also used to analyze sentiments/reviews of employees that were extracted from an online resource ('job planner.co.kr') [10]. In this research, researchers have measured the importance of job satisfaction posted in the reviews related to the company and groups. This research may benefit decision-makers of the respective industries to satisfy their employees by considering the highlighted factors.

Job analytics is gaining importance in IT sectors as well. With time, the requirement of jobs in IT are required on a global level and people prefer to acquire remote jobs to earn high salaries. In this context, job skills reported online by applicants (online) need to be analyzed to filter the skills required by the employer. Using text mining approaches and customized dictionaries, 9000 job postings were analyzed for the position of 'Business analyst' and 'Data analyst' [11]. According to these researchers, the results of these studies mat benefit universities to improve their curriculums as per the skills analyzed through job advertisements by the employers. Similarly, a semi-automated analytical approach was contributed to classifying job skills (IT sector) that intend to help HR managers to devise better recruitment strategies. These researchers scrapped online job posts from 10 different websites to prove the scalability of their approach.

II. Methodology

The comprehensive methodology that has been adopted in this research is explained in this section (as shown in Figure 1). Data was collected and analyzed thoroughly as per the requirement of this research work. After preprocessing the dataset, automated machine learning methods were employed to see the performance of the classification task.

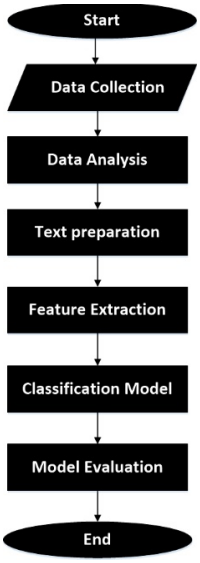

Fig 1. Phases of text processing

a. Data Collection

It was mandatory to collect initial data and information directly from the source to generate a dataset with labels. In this paper, data is collected from www.linkedin.com for finding high-demand job trends. LinkedIn advertised jobs in different fields. For collecting data such as job title, job description, etc. from LinkedIn, we use the Parse Hub web scraping tool as shown in Figure 2. Parse Hub is a free open-source web scrapping tool. It has a good user interface consisting of several selections like selecting only text, images, or certain icons. Parse Hub requires no coding as it consists of simply a drag and drops feature through which data is scraped from a website. Therefore, Parse Hub is applied on the LinkedIn jobs tab. The jobs tab consists of jobs from every location around the world.

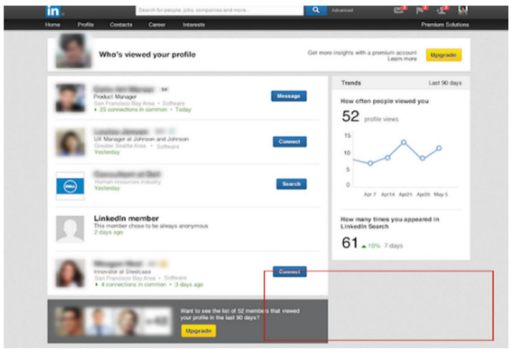

Fig 2. Parse Hub scraping tool

b. Creating an Initial Dataset

The initial dataset was created containing 1200 jobs title/descriptions. This dataset contains the following titles of IT jobs: Android developer, IOS developer, web developer, and front-end developer. During the preparation of the dataset, only a unique job post was included and duplicate data was removed from the dataset. So only beneficial data will remain which is great for further operations.

This research focuses on these 4 categories only. The reason behind this is that the main objective is to classify the text by busing the comparison of two machine learning algorithms so the data classes are not so many users. The data were collected from different countries and did not focus on a specific nation or continent. Because the job description is approximately the same all over the world. So, the data is collected randomly from different countries. The data is not concerned with the junior, senior, or mid-level positions of the above-mentioned categories. So, the job description is general focusing on a different aspect of data. Also, the internship-based position was neglected in the collection of data. The data contains full-time, part-time, and contractual-based jobs. During the preparation of the dataset, only a unique job post was included and duplicate data was removed from the dataset as shown in Figure 3. So only beneficial data will remain which is great for further operations.

Id	Title	FullDescription
30192171	Web Devel	A leading ecommerce agency is looking to hire a Web Developer to join their team of eCommerce Developers
31347600	Web Devel	Web Developer Our client is looking for experienced Web Developers We want people who understand MVC.
46689025	NET WEB [	WEB DEVELOPER/ .NET DEVELOPER ****K TO ****K BASIC PLUS BENEFITS PRESTON 10 MINUTES FROM CIT
55408139	Web Devel	C/.Net Developer Stockton **** **** I am working with a nationally recognised company who are as a resul
55408958	C Senior W	C/.Net WEB Senior Developer Warwick Up to ****k Benefits & Bonus C/ .Net / HTML / CSS / JavaScript /JS [
55409128	Web Devel	Web Developer PHP, MySQL, CSS, HTML A web design and development company based in based in Somerse
55409781	Senior Wel	Senior Web Developer Hands on Development however you must have experience of mentoring junior devel
55409818	Classic ASP	Classic ASP Web Developer JavaScript South London / work from home My client a specialist a software hous
55410140	NET Web [	Fantastic opportunity for an experienced .NET Web Developer to join a global Hitech organisation. The succe

Fig 3. Showing the collected data

Initial dataset based on job titles and job descriptions. After dataset collection applies word cloud using this website www.freewordcloudgenerator.com to generate a cloud of words that frequently appear in the dataset.

c. Data Analysis

Initially, we explore the training dataset and learn about the dataset when applying a text classification problem. Analyzing datasets is a process of identifying and removing inaccurate data. Finding missing data, removing duplicates and unnecessary data, etc. are the pre-processing steps that make the data more accurate. In training, the dataset goes to data analysis using the Python software library pandas. It is an open-source library that provides the best data analysis tool for python programming, and it is easy to use data structures and gives a high rate of performance. The dataset was used consisting of 1200 records with 5000 unique words (approximately). The ratio of classes that have majorities and minorities according to job titles. There are four classes namely web developer, android developer, front-end developer, and iOS developer spread over 58%, 11%, 15%, and 16% respectively.

d. Text Preprocessing

Applying text preprocessing to form an improved version of the dataset facilitates the operation of the machine

learning algorithm. The python library pandas and Natural Language Toolkit (NLTK) were used for analyzing and processing the text [12]. The NLTK is a platform or library that is used to create Python programs that assist in dealing with human language. It contains text processing and machine learning algorithms for Natural Language Processing. By using NLTK the dataset is cleaned by applying preprocessing methods. There are five main techniques of text preprocessing which is used in this research work and applied to the collected dataset:

- *Punctuation Removal*: One of the text techniques is removing punctuations. There is a total of 32 main punctuations which are unnecessary and not useful for our research as well as this punctuation creates problems when applying algorithms to the dataset and accurate results will not be derived. We remove all punctuation from the dataset using this technique.

- *Tokenization*: The second step is text tokenization. In-text tokenization, we split the text string into smaller units or tokens. There are paragraph tokenization or word tokenization. Paragraphs could be split into sentences and sentences into words. In this research paper, we used the word tokenization according to the problem statement. Through these words, the dataset is split into tokens which is useful for the algorithm process.

- *Stop word removal*: It is the third step of text preprocessing, stop words were removed from the dataset which does not add any value to the analysis. These words are commonly used. There is a list of words in the NLTK library that are considered stop words such as I, my, myself, yourself, then, so, and so on. using NLTK library after tokenization of dataset to remove stop words from the dataset.

- *Stemming:* The fourth step is Stemming. Stemming is a process used to extract the base form of a word by removing affixes from the dataset. In this research, we apply to stem to diminish the words into their root form. But there is a problem with this step. Sometimes it stemmed the word into meaningless words and not proper English words. To solve this problem, lemmatization was applied.

- *Lemmatization*: The last step is lemmatization. Lemmatization has a pre-defined dictionary. It also stems from the word but makes sure it does not lose its meaning. After lemmatization, the cleaned dataset became ready to apply the machine learning algorithm.

The collected IT job data are now cleaned by removing certain special characters, numbers, whitespace characters, and stop words by applying text preprocessing. Now dataset is prepared for vectorization because vectorization is most important before applying a machine learning algorithm. The word cloud is also changed after text preprocessing because many unnecessary words were removed through text preprocessing.

e. Feature Extraction

Feature extraction is a process of converting the dataset i.e., the input dataset into numbers. Since machine learning algorithms receive only numbers therefore features should be extracted to make the vector. A vector is a set of numbers related to words in feature extraction the columns are words and the rows are the occurrence of words in that row. E.g., the word is tested so the vector shows how many times these words appear in each sentence suppose there are 3000 sentences so how many times does the word appear in each sentence. To apply machine learning algorithms, it is mandatory to convert the text string into vector representations since machine learning algorithms work in a numeric feature. Input is considered as a two-dimensional array where columns are features and rows are instances. There are many feature extraction methods available.

In this research, the most popular feature extraction methods are used. Term Frequency-Inverse Document Frequency (TF-IDF) and count vectorization are applied for feature extraction using the NLTK library. TF-IDF method is a numerical statistic whose aim is to find the importance of a word in a collection of datasets. The second technique count vectorization used to convert a collection of text data into a token count. A word that seems a minimum of 5 times in the entire dataset is considered a feature. A total of 100 features were extracted by using the TF-IDF method as shown in Figure 4.

backend	bug	codeignit	coffee	consultan	data	database	debuggin	developer	drupal	frontend	github	graphics	html
0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	2	0	0	0	0	1
0	0	0	0	0	0	0	0	25	0	0	0	0	2
0	0	0	0	0	0	0	0	2	0	0	0	0	1
0	0	0	1	0	0	0	0	3	0	0	1	0	1
0	0	0	0	0	0	0	0	2	0	0	0	0	2
0	0	0	0	0	0	0	0	2	0	0	0	0	0
0	0	0	0	0	0	0	0	2	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	1
0	0	0	0	0	0	0	0	4	1	0	0	0	2
0	0	0	0	0	0	0	1	4	0	0	0	0	0

Fig 4. Dataset after vectorization

f. Creating a Text Classification Model

After text preprocessing, the dataset is cleaned from numbers, stop words, punctuation, and unnecessary words. Feature extraction extract features from the processed dataset. The next step is to build a text classification model. The text classification model is based on the training dataset. The model is used to predict the categories.

We perform text classification using python language, as it is an open-source resource that could help to implement text classification without any problem. ScikitLearn open-source library is used in this research. This library may contain many machine learning algorithms such as Multinomial naïve Bayes and Support Vector Machine [13]. Multi-class classification machine learning algorithms are based on the following steps: importing libraries, fetching the dataset into the data frame as shown in Figure 5, features extracting, train-test dataset splitting, training the model, and calculating the model result using the appropriate metric. Dataset is divided into training and test datasets using Scikit-learn, train, test, and split function. In this experiment, the

sample column is the job description. In Figure 5, 70% of the data is used as training data and 30% is used for the testing purpose. The training dataset is used to create the model and the test dataset is used to validate the trained model.

Id	Title	FullDescription
30192171	Web Deve	A leading ecommerce agency is looking to hire a Web Developer to join their team of eCommerce Developer
31347600	Web Deve	Web Developer Our client is looking for experienced Web Developers We want people who understand MV
46689025	NET WEB I	WEB DEVELOPER/ .NET DEVELOPER ****K TO ****K BASIC PLUS BENEFITS PRESTON 10 MINUTES FROM CITY CEN
55408139	Web Deve	C/.Net Developer Stockton **** **** I am working with a nationally recognised company who are as a result
55408958	C Senior V	C/.Net WEB Senior Developer Warwick Up to ****k Benefits & Bonus C/ .Net / HTML / CSS / JavaScript /JS D
55409128	Web Deve	Web Developer PHP, MySQL, CSS, HTML A web design and development company based in based in Somers
55409781	Senior We	Senior Web Developer Hands on Development however you must have experience of mentoring junior dev
55409818	Classic AS	Classic ASP Web Developer JavaScript South London / work from home My client a specialist a software hou
55410140	NET Web I	Fantastic opportunity for an experienced .NET Web Developer to join a global Hitech organisation. The succe
55410405	PHP Web	Description: PHP Developer / Web Developer Bournemouth PHP, MySQL, JavaScript, HTML/CSS Our multiaw

Fig 5. Dataset before text preprocessing.

g. Model Evaluation

After implementing feature engineering, selection, training the model, and getting some results in form of probability. The next task is to validate how accurate the model is based on some metric using the test data created earlier. The performance of the model is explained by the matrix.

For multi-class problems, the performance of the classifier was defined by a confusion matrix related to the classifier. The confusion matrix is a table where each cell (x, y) describes how often label y was predicted when the correct label was x. So, the diagonal line indicates the correct predicted labels, and the diagonal shows error.

There are four classes created from a dataset i.e., android, IOS, web, and frontend. Python was used to calculate parameters or performance of classes through the calculation performance metrics such as accuracy, precision, recall, and F1-score. Scikit-Learn metrics and Python Seaborn 0.10.1 visualization library were used in this calculation.

All parameters for them were calculated in Python, using Scikit-Learn metrics and a Python Seaborn 0.10.1 visualization library. There are two target values positive and negative. The row represents the predicted values and the column represents the actual values.

h. Evaluation Through Confusion Matrix

The confusion matrix for the linear support vector machine is shown in Figure 6 and the multinomial naïve Bayes classifier is shown in Figure 7.

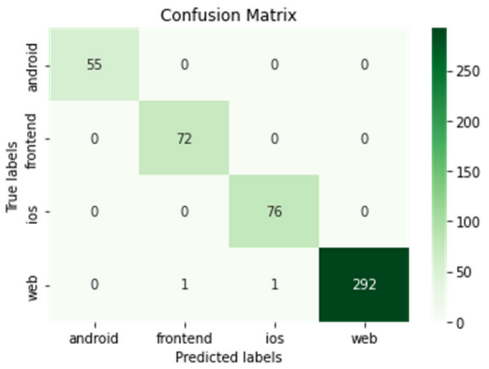

Fig 6. A 4x4 confusion matrix for the Linear SVC

The confusion matrix in Figure 6 defines the performance of the classification of the linear support vector machine algorithm. They form a great accuracy as compared to the multinomial naïve Bayes classifier.

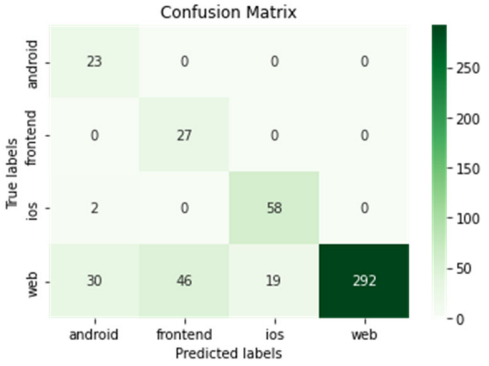

Fig 7. A 4x4 confusion matrix for the MultinomialNB

The confusion matrix in Figure 7 defines the performance of the classification of the multinomial Naïve Bayes machine algorithm. They form a low accuracy as compared to linear support vector machine classifiers. After this calculate the result by applying formulas of (precision, recall, F1-score, and support).

III. Results

Accuracy is commonly used as an evaluation metric to analyze performance [14]. Many techniques are implemented to enhance the quality of text translation [15, 16] and data mining [17]. A classification report describes classification metrics (Precision, Recall, F1-score, and support) in form of a text report. It provides an overall performance of a machine learning trained model. Figure 6 and Figure 7 show a classification report of Multinomial Naïve Bayes and Linear Support vector machine algorithm.

TABLE I. CLASSIFICATION REPORT OF NAÏVE BAYES ALGORITHM

Categories	Precision	Recall	F1-score	Support
Android Developer	1.00	0.42	0.59	55
Front-end web Developer	1.00	0.37	0.54	73
IOS Developer	0.97	0.75	0.85	77
Web Developer	0.75	1.00	0.86	292
Accuracy			0.80	497
Macro Average	0.93	0.64	0.71	497
Weighted Average	0.85	0.80	0.78	497

Table I shows the accuracy, precision, recall, F1-score, and support of the Multinomial Naïve Bayes algorithm. According to the results obtained, the android developer and front-end web developer are less demanding jobs than iOS and web developers which means that the classifier has predicted the positive class as a negative class due to which the ratio is low. Therefore, the overall accuracy becomes 80% due to the wrong prediction of those classes. The reason for average accuracy is due to the negative prediction of classes.

TABLE II. CLASSIFICATION REPORT OF LINEAR SUPPORT VECTOR ALGORITHM

Categories	Precision	Recall	F1-score	Support
Android Developer	1.00	1.00	1.00	55
Front-end web Developer	1.00	0.99	0.99	73
IOS Developer	1.00	0.99	0.99	77
Web Developer	0.99	1.00	1.00	292
Accuracy			1.00	497
Macro Average	1.00	0.99	1.00	497
Weighted Average	1.00	1.00	1.00	497

Table II shows the accuracy, precision, recall, F1-score, and support of the Support Vector Machine algorithm. According to the obtained result, the algorithm has correctly predicted the majority of classes as positive classes. The Support Vector Machine in this case produces good results as compared to Multinomial Naïve Bayes.

Accuracy alone is not enough for the evaluation of both classifiers. Since the model is a multiclass classification model and the dataset has an unequal number of observations for each class. Therefore, in this case, observing the accuracy can only lead to a wrong estimation of the result. The classification report shows precision and recalls an additional feature and it helps to identify which model is more perfect or accurate.

In the case of high precision and recall it is an indicator that the classifier returns the correct result (precision) and returns most of all positive results (recall).

IV. Conclusion

This research paper defines the use of Multinomial Naïve Bayes and Linear Support vector machine learning algorithm for text classification on a dataset. In this paper, we create a trained multilingual dataset and data that contains two field job titles and job descriptions after applying text preprocessing. We remove many unnecessary things and generate a list of stop words.

Performance results showed the beneficial use of supervised machine learning algorithms for the classification of jobs according to their categories. By the reference to classification report shown in Figures 5 and 6, it is clear that the Linear Support vector machine algorithm gives better results in comparison with Multinomial Naïve Bayes. The precision and recall measures along with the accuracy are evidence that the Linear Support vector machine algorithm can handle the above dataset well. Although Multinomial Naïve Bayes also provides average performance.

In the future, we aim to perform the knowledge-based classification on a big dataset as we did in this research paper. We will also use dataset occupation codes which are not included in this the dataset that is used in this research work. To achieve this goal, we will work with deep learning algorithms and can create a new or trained dataset to accomplish this goal. Moreover, we can apply another machine learning algorithm for text classification to produce statistical results and explore the more possible use of machine learning algorithms to classify the textual data collected by regular statistical surveys.

References

[1] Boselli R, Cesarini M, Mercorio F, Mezzanzanica M. Classifying online job advertisements through machine learning. Future Generation Computer Systems. 2018 Sep 1;86:319-28..

[2] Keerthi Kumar HM, Harish BS. Classification of short text using various preprocessing techniques: An empirical evaluation. InRecent findings in intelligent computing techniques 2018 (pp. 19-30). Springer, Singapore.

[3] Gurcan F, Cagiltay NE. Big data software engineering: Analysis of knowledge domains and skill sets using LDA-based topic modeling. IEEE access. 2019 Jun 20;7:82541-52.
[4] Hartmann J, Huppertz J, Schamp C, Heitmann M. Comparing automated text classification methods. International Journal of Research in Marketing. 2019 Mar 1;36(1):20-38.
[5] Cvijetić B, Radivojević Z. Application of machine learning in the process of classification of advertised jobs. International Journal of Electrical Engineering and Computing. 2020 Dec 22;4(2):93-100.
[6] Bansal S, Srivastava A, Arora A. Topic modeling driven content based jobs recommendation engine for recruitment industry. Procedia computer science. 2017 Jan 1;122:865-72.
[7] Pejić Bach M, Krstić Ž, Seljan S, Turulja L. Text mining for big data analysis in financial sector: A literature review. Sustainability. 2019 Feb 28;11(5):1277.
[8] Jung Y, Suh Y. Mining the voice of employees: A text mining approach to identifying and analyzing job satisfaction factors from online employee reviews. Decision Support Systems. 2019 Aug 1;123:113074.
[9] De Mauro A, Greco M, Grimaldi M, Ritala P. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. Information Processing & Management. 2018 Sep 1;54(5):807-17.
[10] Dong T, Triche J. A longitudinal analysis of job skills for entry-level data analysts. Journal of Information Systems Education. 2020 Dec 1;31(4):312.
[11] Amado A, Cortez P, Rita P, Moro S. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics. 2018 Jan 1;24(1):1-7.
[12] Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc."; 2009 Jun 12.
[13] Ramachandran D, Parvathi R. Analysis of twitter specific preprocessing technique for tweets. Procedia Computer Science. 2019 Jan 1;165:245-51.
[14] Hussain A, Ali G, Akhtar F, Khand ZH, Ali A. Design and Analysis of News Category Predictor. Engineering, Technology & Applied Science Research. 2020 Oct 26;10(5):6380-5.
[15] Chopra D, Joshi N, Mathur I. Improving Translation Quality By Using Ensemble Approach. Engineering, Technology & Applied Science Research. 2018 Dec 22;8(6):3512-4.
[16] Khan U, Khan K, Hassan F, Siddiqui A, Afaq M. Towards achieving machine comprehension using deep learning on non-GPU machines. Engineering, Technology & Applied Science Research. 2019 Aug 10;9(4):4423-7.
[17] Kazi AK, Nazir W, Baig MA and Khan S. Breast Cancer Prediction Using Data Mining Classification Techniques. International Journal of Computer Science And Network (IJCSN). 2022 Sep;22(9).