

Resume Selection using Machine Learning Technique

By

Rakin Mohammad Sifullah

Abstract:

Finding suitable candidates for an open role could be a daunting task, especially when there are many applicants. It can impede team progress in getting the right person at the right time. An automated way of “Resume Classification and Matching” could really ease the tedious process of fair screening and shortlisting, it would certainly expedite the candidate selection and decision-making process.

Talent acquisition is an important, complex, and time-consuming function within Human Resources (HR). The sheer scale of the Bangladeshi market is overwhelming. Not only is there a staggering one million people coming into the job market every month, but there is also a huge turnover. Clearly, this is an extremely liquid, massive market but one that also has many frustrating inefficiencies. The most challenging part is the lack of a standard structure and format for resumes which makes a short listing of desired profiles for required roles very tedious and time-consuming. Effective screening of resumes requires domain knowledge, to be able to understand the relevance and applicability of a profile for the job role. With a huge number of different job roles existing today along with the typically large number of applications received, short-listing poses a challenge for the human resource department. This is only further worsened by the lack of diverse skills and domain knowledge within the HR department, required for effectiveness. screening. Being able to weed out non-relevant profiles as early as possible in the pipeline results in cost savings, both in terms of time as well as money.

Today the industry faces three major challenges:

- Separating the right candidates from the pack - India is a huge job market and with millions seeking jobs; it is humanly impossible to screen the CVs and find the right match. This makes the whole hiring process slow and inefficient costing resources the companies.
- Making sense of candidate CVs - The second challenge is posed by the fact that the CVs in the market are not standard practically every resume in the market has a different structure and format. HR has to manually go through the CVs to find the right match for the job description. This is resource intensive and prone to error whereby the right candidate for the job might get missed in the process.
- Knowing that candidates can do the job before you hire them -The third and major challenge is mapping the CV to the job description to understand if the candidate would be able to do the job for which she is being hired.

To overcome the mentioned issues in the resume short-listing process, an automated Machine Learning based model must be needed. The model takes the features extracted from the candidate's resume as input and finds their categories, further based on the required job description the categorized resume is mapped and recommends the most suitable candidate's profile to HR. The main contributions are listed below:

1. Developed an automated resume recommendation system.
2. Machine learning-based classification techniques with similarity functions are used to find the most relevant resume.

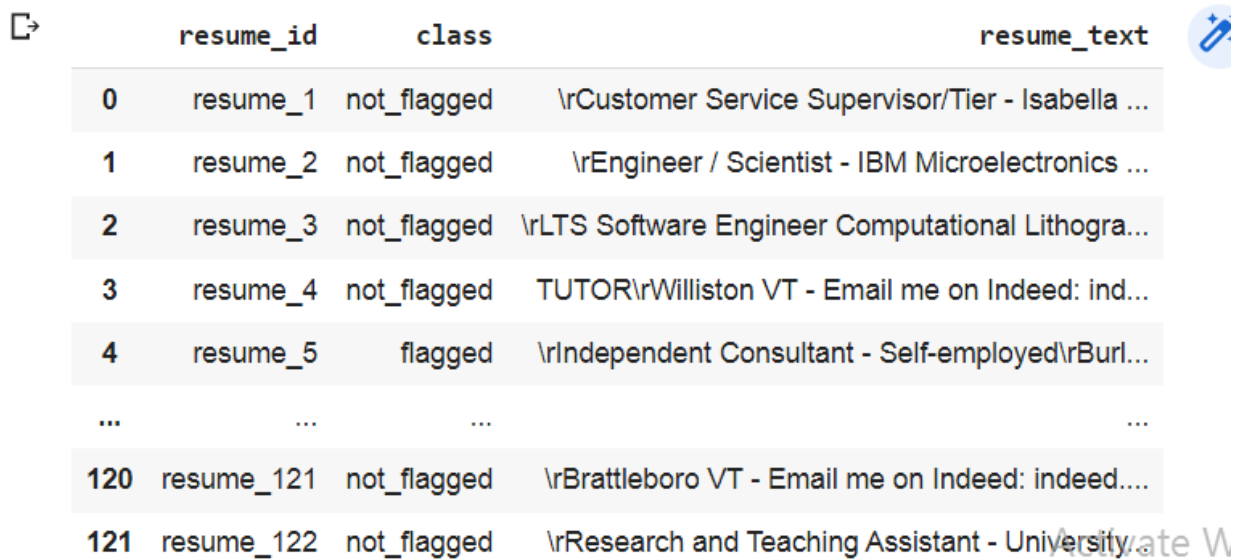
Methodology:

The aim of this work is to find the right candidate's resume from the pool of resumes. To achieve this objective, we have developed a machine learning-based solution. The complete framework for the proposed model is shown in Figure 2. The proposed model worked mainly in two steps:

- Prepare and
- Deploy and Inference.

Dataset Description:

The data was downloaded from the online portal(s). The data is in Excel format, with three columns resume_id, class, and resume_text. Resume_id - The sequence number of the resume, Class - Industry sector to which the resume belongs to, and Resume_text - The complete CV of the candidate. The number of instances for the different domains can be seen in Figure 1.



	resume_id	class	resume_text
0	resume_1	not_flagged	\rCustomer Service Supervisor/Tier - Isabella ...
1	resume_2	not_flagged	\rEngineer / Scientist - IBM Microelectronics ...
2	resume_3	not_flagged	\rLTS Software Engineer Computational Lithogra...
3	resume_4	not_flagged	TUTOR\rWilliston VT - Email me on Indeed: ind...
4	resume_5	flagged	\rIndependent Consultant - Self-employed\rBurl...
...
120	resume_121	not_flagged	\rBrattleboro VT - Email me on Indeed: indeed....
121	resume_122	not_flagged	\rResearch and Teaching Assistant - University

Figure-1: Dataset

Vectorizer:

For vectorize here, **CounterVectorizer** has been used.

CountVectorizer is a great tool provided by the scikit-learn library in Python. It is used to transform a given text into a vector on the basis of the frequency (count) of each word that occurs in the entire text.

Model:

For this Work, **Multinomial Naive Bayes** has been used.

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.

Preprocessing:


In this process, the CVs being provided as input would be cleansed to remove special or any junk characters that are there in the CVs. In cleaning, all special characters, numbers, and single-letter words are removed. We got the clean dataset after these steps having no special characters, numbers, or single letter words. The dataset is split into the tokens using the NLTK tokenizes. Further, the preprocessing steps are applied to a tokenized dataset such as stop word removal, stemming, and lemmatization. The raw CV file was imported and the data in the resume field was cleansed to remove the numbers and the extra spaces in the date.

Steps of the work :


1. Importing Libraries

- INSTALLING NLTK, GENSIM AND WORDCLOUD
- pandas, numpy, matplotlib, seaborn, nltk, gensim, sklearn, wordcloud

2. Loading Dataset



	resume_id	class	resume_text
0	resume_1	not_flagged	\rCustomer Service Supervisor/Tier - Isabella ...
1	resume_2	not_flagged	\rEngineer / Scientist - IBM Microelectronics ...
2	resume_3	not_flagged	\rLTS Software Engineer Computational Lithogra...
3	resume_4	not_flagged	TUTOR\rWilliston VT - Email me on Indeed: ind...
4	resume_5	flagged	\rIndependent Consultant - Self-employed\rBurl...
...
120	resume_121	not_flagged	\rBrattleboro VT - Email me on Indeed: indeed....
121	resume_122	not_flagged	\rResearch and Teaching Assistant - University...



3. Performing Exploratory Data Analysis

```
not_flagged    92
flagged        33
Name: class, dtype: int64
```

4. Performing Data Cleaning

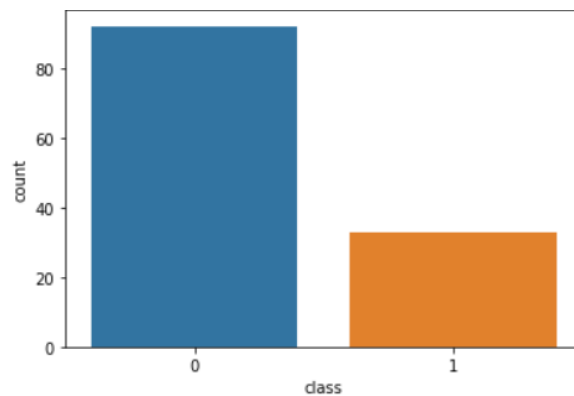
- REMOVING UNNECESSARY WORDS FROM DATASET

[8]

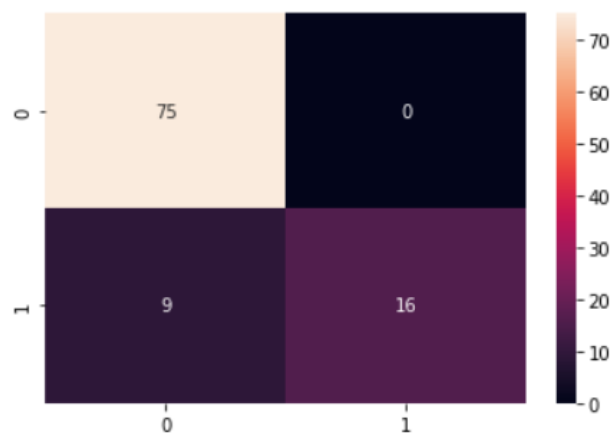
	resume_text	class
0	Customer Service Supervisor/Tier - Isabella Ca...	0
1	Engineer / Scientist - IBM Microelectronics Di...	0
2	LTS Software Engineer Computational Lithograph...	0
3	TUTORWilliston VT - Email me on Indeed: indee...	0
4	Independent Consultant - Self-employedBurlingt...	1
...
120	Brattleboro VT - Email me on Indeed: indeed.co...	0
121	Research and Teaching Assistant - University o...	0
122	Medical Coder - Highly Skilled - Entry LevelSu...	0
123	Waterbury VT - Email me on Indeed: indeed.com/...	1

5. Visualizing Cleaned Dataset

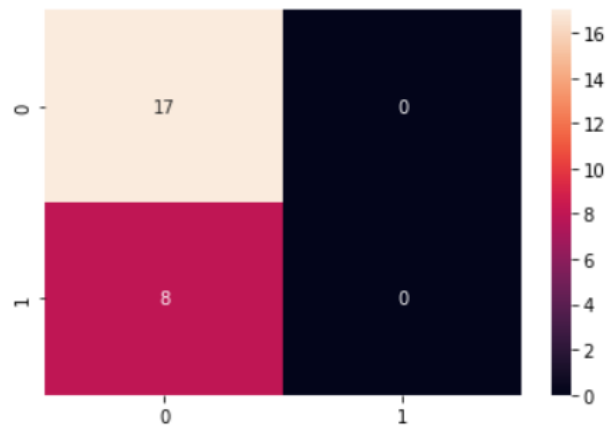
- PLOTTING COUNTS OF SAMPLE LABELLED AS 1 AND 0



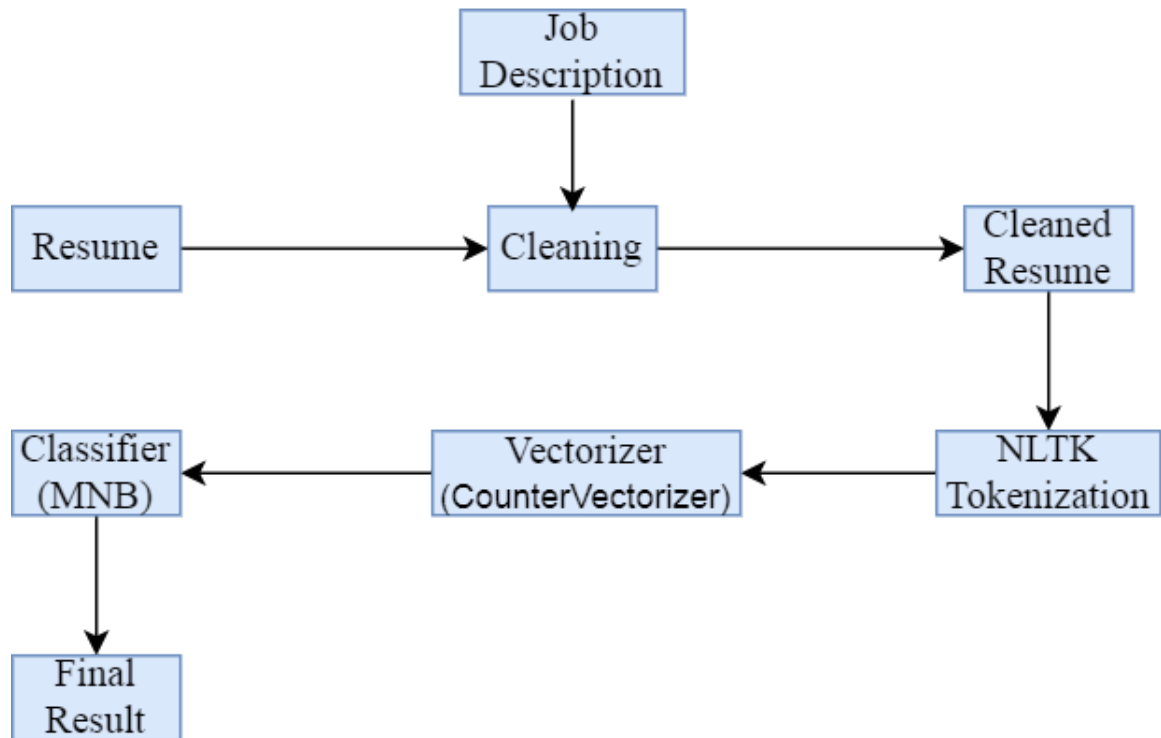
- PLOTTING THE WORDCLOUD
- CONVERTING SENTENCES INTO TOKENIZED FORMS AND THEN CONVERTING TO NUMERICAL VALUES IN ORDER FOR THE MODEL TO TRAIN
- PLOTTING CONFUSION MATRIX: FOR TRAINING DATA



- WE CAN SEE OUR MODEL PERFORMED REALLY WELL ON TRAINING DATA: IT CLASSIFIED ALL OF THE POINTS CORRECTLY



A complete framework of the proposed model:



Results:

Final Accuracy: 0.68