Resume Analysis using Natural Language Processing

For Data Analytics/ Data Science Job Roles

Akanshya Meher
Data Analytics
University of Central Florida
Orlando FL USA
akanshya.meher@knights.ucf.edu

Puja Ramesh Sonawane
Data Analytics
University of Central Florida
Orlando FL USA
puja.sonawane@knights.ucf.edu

Jahnavi Sandhya Lanka
Data Analytics
University of Central Florida
Orlando FL USA
jahnavisandhyalanka@knights.ucf.edu

ABSTRACT

The paper comprises of measures to aid employees and candidates to analyze resumes just like how a real-world ATS (Application Tracking Systems) analyzes it, in order to help candidates, improve the skills required in the job descripting and meanwhile helping employers select the best candidates according to their standards and preferences. The paper has used NLP (Natural Language Processing) techniques for text summarization, pattern matching and skill wise classification.

INTRODUCTION

About 90% of the leading industries' Human Resource departments has used ATS (Application Tracking System) to filter suitable candidates automatically from large application pool before a resume is analyzed by a human representative, leading to less manpower and more efficiency. Considering this, it has been imperative for candidates to know the working understanding of ATS in order to be considered any job role in the competitive market.

The system introduced in this paper not only helps candidates to improve upon the skills required in a job but also to any employer looking for the right candidates for respective job roles. This problem was addressed by presenting similar analysis in terms of visualizations and similarity/dissimilarity measures to predict accuracies of the outputs.

It has been proven evident that during a pandemic, recession where there is a huge dip present in the job market, it calls for an ideal system where it has become more crucial than ever to prepare a resume that matches the job description for respective positions and to be considered in the application process by interview rounds.

ACKNOWLEDGMENTS

Our greatest thanks to the students of Advanced Text Mining class for helping use provide their accurate and reliable sample resumes for testing to achieve precise results. In addition, we would like to give extra thanks to our Professor, Dr. Awrad Mohammad Ali who helped us understand the concepts and techniques of the subject without which the project would never have succeeded and helped many of us trying to land a successful job after the master's program.

KEYWORDS

ATS, NLP, Text Summarization, Text Classification, Phrase Matcher, Python, Similarity, Dissimilarity, Spell Checker.

1. RELATED WORK

To date, there has been many existing papers using methods of text summarization and classification using NLP due to its popularity and effectiveness. In order to understand our work better and how our work differs from the existing methods, some relevant research papers which aligns with a common interest will be introduced and discussed further.

1.1 Resumatcher

Resumatcher: A personalized resume-matching job system (Shiqiang Guo, 2015) proposed a system to help job seekers seeking jobs easily by creating a finite state transducer based on data extraction library to extract models from job descriptions and resumes which were mainly unstructured. The paper introduced a new statistical-based ontology similarity measure in order to compare and contrast the models of resume and jobs; additionally, measured the similarities among technical terms.

The idea was to return jobs based on appropriateness as compared to traditional job searching websites. In terms of verifying and the validity of the models, the author computed NDCG (Normalized Discounted Cumulative Gain) which is a ratio of DCG (Discounted Cumulative Gain) and IDCG (Ideal Discounted Cumulative Gain) as well as precision meanwhile also comparing three existing models that is Okapi BM25, TF-IDF and Kullback-Leibler divergence using the live results from Indeed.com.

1.2 Resume Ranking

Resume Ranking using NLP and Machine Learning (Zubeda, Shaheen, Godavari and Naseem, 2016) has used a ranking algorithm to rank resumes of any format mapping with the constraints given by a company. Moreover, the paper has used NLP and ML techniques to read applicants various social media profiles like Github, LinkedIn, etc. despite the existing resumes in order to retrieve accurate information of applicants.

1.3 Limitations

The existing papers do give a high scope to reach a common solution and helping candidates as well as employers to select to the best. Even though, our paper uses some common techniques used by the above, the key underlying difference is that our paper uses the existing NLP techniques using CountVectorizer for matching and summarization and verifying the models using well defined similarity measure of Cosine and dissimilarity measure of Jaccard.

Additionally, the missing keywords which are not present in the resume and are a necessity in the job description are pointed out to the candidate which could be a reason for a reject on the application.

Both the papers have used different ways of input and our system used PDF and Word files for inputs which are the highly common resume files used by candidates. The methods to extract the given files and evaluation of models will be discussed in the upcoming sections.

2. METHOD

Built a model that scans a group of users' resume and highlights skills/technologies using bar chart and pie chart that assists employers to choose a person with the highest skill they require through detailed visualizations. The model is beneficial for applicants and employers to check similarity among job description and resumes for Data Analytics / Data Science job roles.

Language: Python

Techniques: Pattern Matching, Text Summarization, Skill wise Classification and Spell Checker.

Algorithm: Count Vectorizer

Libraries: numpy, pandas, re, spaCy, docx2txt, pydf2, nltk, textract, genism and CountVectorizer.

Evaluation metric: Cosine Similarity and Jaccard Similarity.

The paper has used libraries like Phrase Matcher from spaCy to match phrases from the resume with the keyword dictioneries, pyPDF2 and docx2txt to extract information from resumes that are in PDF and Word formats respectively, genism for similarity retrieval for summarization.

3. EXPERIMENTS

Six experiments were conducted in the paper which comprises of: Count Vectorizer, Spell Checker, Text Summarization, Similarity Scores, Missing Keywords and Skill Wise Classification and Comparison. These experiments would be further discussed in detail in the upcoming sections.

3.1 Count Vectorizer and Cosine Similarity

The resumes are first converted from an unstructured data source to a structured textual format and then feature extraction is applied to it.

Feature extraction was achieved using the concept of count vectorizer which uses predictive modeling for text data by either parsing through the data or removing words which is also known as tokenization. CountVectorizer using this idea converts the text data into vector of token counts which helps in the preprocessing of text data making it flexible at the same time.

The following is an example on how CountVectorizer works:
Text = The quick brown fox jumps over the lazy dog
Data = ['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy' ,'dog']
Count Vector =

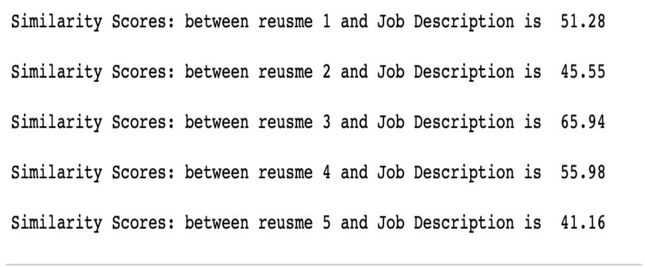
The	quick	brown	fox	jumps	over	the	lazy	dog
2	1	1	1	1	1	1	1	1

Using the underlying notion of CountVectorizer, it was applied on the converted input of resumes. After this stage, cosine similarity is applied which is a similarity of two documents regardless of their size. In this paper, it measures the cosine angle between the two count vectors achieved in the previous stage. The formula for it is as follows:

$$Cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \, \|\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2} \sqrt{\sum_1^n b_i^2}} \qquad (1)$$

where, $\vec{a} \cdot \vec{b} = \sum_1^n a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ is the dot product of the two vectors.

Computing the above formula in python on a sample of 5 resumes and job description for entry level Data Analyst available on Indeed.com the following results were achieved:



```
Similarity Scores: between reusme 1 and Job Description is  51.28

Similarity Scores: between reusme 2 and Job Description is  45.55

Similarity Scores: between reusme 3 and Job Description is  65.94

Similarity Scores: between reusme 4 and Job Description is  55.98

Similarity Scores: between reusme 5 and Job Description is  41.16
```

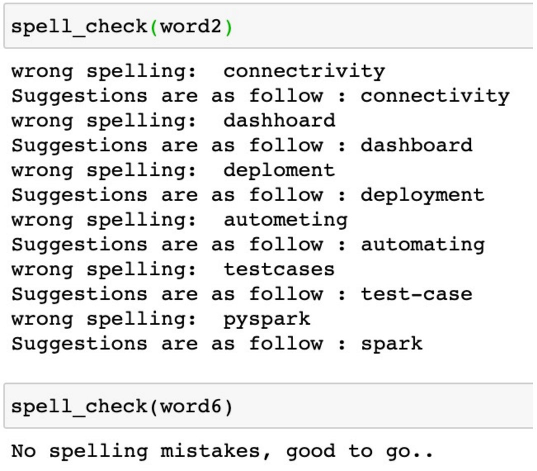
Figure 1: Cosine similarity scores among five different data analyst candidates and entry-level data analyst job description.

Observation from the figure above suggests that candidate 3 has the maximum similarity score, that is, 65.94 with respect to job description and therefore has a higher chance of getting selected for the job as opposed to candidate 5 which has a considerably lower similarity score, that is, 41.16 with respect to job description and is likely to not be shortlisted for the job compared to all the other four candidates. These results can be insightful for the candidates when applying for a job considering they might be lacking in the crucial required skills which a prerequisite in the job description and likewise the candidate can improve upon it to be considered for the job role.

3.2 Spell Checker

One very common mistake any candidate can make is using incorrect words in their resume unknowingly. However, this might not seem like a huge error but sometimes the ATS disregards any word that might be misspelled, and only consider similarity scores with the correct keywords.

For this reason, spell checker was implemented in this paper with a built-in function in python to help candidates not make these mistakes and even if they do, it can be corrected. The following was the output of a resume which identified the wrong spelling and suggested an alternative indeed.



```
spell_check(word2)
```

```
wrong spelling:  connectrivity
Suggestions are as follow : connectivity
wrong spelling:  dashhoard
Suggestions are as follow : dashboard
wrong spelling:  deploment
Suggestions are as follow : deployment
wrong spelling:  autometing
Suggestions are as follow : automating
wrong spelling:  testcases
Suggestions are as follow : test-case
wrong spelling:  pyspark
Suggestions are as follow : spark
```

```
spell_check(word6)
```

```
No spelling mistakes, good to go..
```

Figure 2: Spell check for the words – connectivity, dashboard, deployment, automating, test-case and spark.

The function used in the above program identified the incorrect word used in a resume and corrected them, for example, connectrivity was changed to connectivity, autometing was changed to automating, deploment was changed to deployment and so on.

If the resume had all correct words, then the program gives a go sign stating no spelling mistakes were observed which in turn means the resume is ready for further analysis for similarity and classification which will be discussed in the next section.

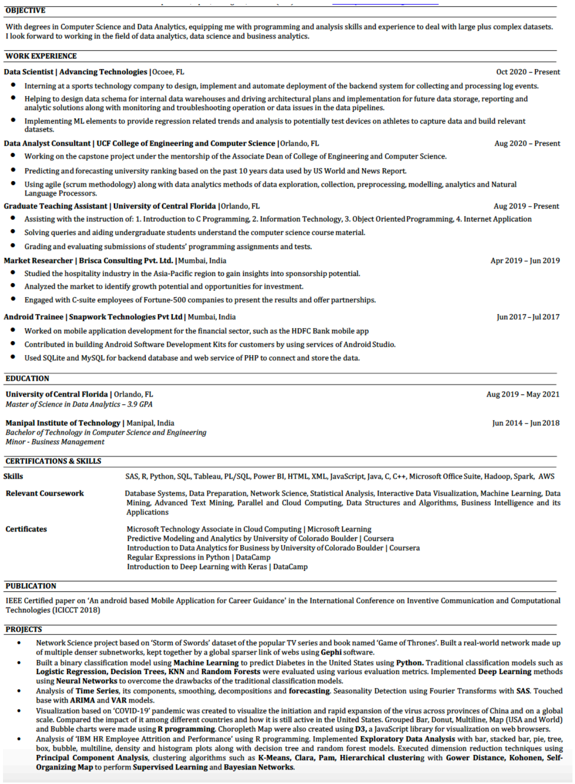
3.3 Text Summarization

Text Summarization is one of the most interesting topics in NLP with the goal of reducing reading time and increasing computational speed for modeling and accuracy. It is a process of getting meaningful text with semantics from a large document. It can be applied to articles, journals, research papers like this one and even a resume.

The reason why this paper has implemented text summarization is because sometimes resumes can be of more than one page or a lot of information can be present in a single page. It might not be ideal to read the entire document and hence an ATS system use a shortened version of the existing resume in order to find similarities among the resume and the job description.

Keeping this in mind, this paper has summarized about 10 sample resumes but has not summarized the job description. The cause behind this decision was taken because job description in itself is likely to be quite specific and missing important keywords would result in inaccurate results. However, this would not be the case with a resume since there might be existing information which is not a necessity for pattern matching and computing similarity scores.

The following is a resume of one page which has been summarized with a ratio of 10%.



OBJECTIVE

With degrees in Computer Science and Data Analytics, equipping me with programming and analysis skills and experience to deal with large plus complex datasets. I look forward to working in the field of data analytics, data science and business analytics.

WORK EXPERIENCE

Data Scientist | Advancing Technologies |Ocoee, FL Oct 2020 – Present

- Interning at a sports technology company to design, implement and automate deployment of the backend system for collecting and processing log events.
- Helping to design data schema for internal data warehouses and driving architectural plans and implementation for future data storage, reporting and analytic solutions along with monitoring and troubleshooting operation or data issues in the data pipelines.
- Implementing ML elements to provide regression related trends and analysis to potentially test devices on athletes to capture data and build relevant datasets.

Data Analyst Consultant | UCF College of Engineering and Computer Science |Orlando, FL Aug 2020 – Present

- Working on the capstone project under the mentorship of the Associate Dean of College of Engineering and Computer Science.
- Predicting and forecasting university ranking based on the past 10 years data used by US World and News Report.
- Using agile (scrum methodology) along with data analytics methods of data exploration, collection, preprocessing, modelling, analytics and Natural Language Processors.

Graduate Teaching Assistant | University of Central Florida |Orlando, FL Aug 2019 – Present

- Assisting with the instruction of: 1. Introduction to C Programming, 2. Information Technology, 3. Object Oriented Programming, 4. Internet Application
- Solving queries and aiding undergraduate students understand the computer science course material.
- Grading and evaluating submissions of students' programming assignments and tests.

Market Researcher | Brisca Consulting Pvt. Ltd. |Mumbai, India Apr 2019 – Jun 2019

- Studied the hospitality industry in the Asia-Pacific region to gain insights into sponsorship potential.
- Analyzed the market to identify growth potential and opportunities for investment.
- Engaged with C-suite employees of Fortune-500 companies to present the results and offer partnerships.

Android Trainee | Snapwork Technologies Pvt Ltd | Mumbai, India Jun 2017 – Jul 2017

- Worked on mobile application development for the financial sector, such as the HDFC Bank mobile app
- Contributed in building Android Software Development Kits for customers by using services of Android Studio.
- Used SQLite and MySQL for backend database and web service of PHP to connect and store the data.

EDUCATION

University of Central Florida | Orlando, FL Aug 2019 – May 2021
Master of Science in Data Analytics – 3.9 GPA

Manipal Institute of Technology | Manipal, India Jun 2014 – Jun 2018
Bachelor of Technology in Computer Science and Engineering
Minor - Business Management

CERTIFICATIONS & SKILLS

Skills	SAS, R, Python, SQL, Tableau, PL/SQL, Power BI, HTML, XML, JavaScript, Java, C, C++, Microsoft Office Suite, Hadoop, Spark, AWS
Relevant Coursework	Database Systems, Data Preparation, Network Science, Statistical Analysis, Interactive Data Visualization, Machine Learning, Data Mining, Advanced Text Mining, Parallel and Cloud Computing, Data Structures and Algorithms, Business Intelligence and its Applications
Certificates	Microsoft Technology Associate in Cloud Computing | Microsoft Learning Predictive Modeling and Analytics by University of Colorado Boulder | Coursera Introduction to Data Analytics for Business by University of Colorado Boulder | Coursera Regular Expressions in Python | DataCamp Introduction to Deep Learning with Keras | DataCamp

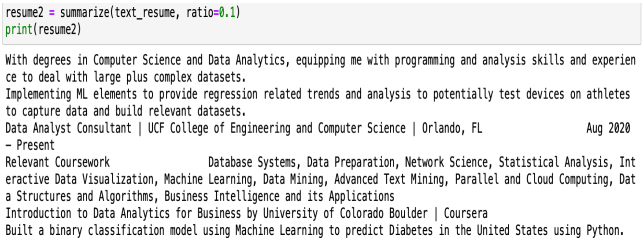
PUBLICATION

IEEE Certified paper on 'An android based Mobile Application for Career Guidance' in the International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)

PROJECTS

- Network Science project based on 'Storm of Swords' dataset of the popular TV series and book named 'Game of Thrones'. Built a real-world network made up of multiple denser subnetworks, kept together by a global sparser link of webs using **Gephi** software.
- Built a binary classification model using **Machine Learning** to predict Diabetes in the United States using **Python.** Traditional classification models such as **Logistic Regression, Decision Trees, KNN** and **Random Forests** were evaluated using various evaluation metrics. Implemented **Deep Learning** methods using **Neural Networks** to overcome the drawbacks of the traditional classification models.
- Analysis of **Time Series**, its components, smoothing, decompositions and **forecasting**. Seasonality Detection using Fourier Transforms with **SAS**. Touched base with **ARIMA** and **VAR** models.
- Visualization based on 'COVID-19' pandemic was created to visualize the initiation and rapid expansion of the virus across provinces of China and on a global scale. Compared the impact of it among different countries and how it is still active in the United States. Grouped Bar, Donut, Multiline, Map (USA and World) and Bubble charts were made using **R programming**. Choropleth Map were also created using **D3,** a JavaScript library for visualization on web browsers.
- Analysis of 'IBM HR Employee Attrition and Performance' using R programming. Implemented **Exploratory Data Analysis** with bar, stacked bar, pie, tree, box, bubble, multiline, density and histogram plots along with decision tree and random forest models. Executed dimension reduction techniques using **Principal Component Analysis**, clustering algorithms such as **K-Means, Clara, Pam, Hierarchical clustering** with **Gower Distance, Kohonen, Self-Organizing Map** to perform **Supervised Learning** and **Bayesian Networks**.

Figure 3: Original resume taken from the user.



```
resume2 = summarize(text_resume, ratio=0.1)
print(resume2)
```

```
With degrees in Computer Science and Data Analytics, equipping me with programming and analysis skills and experien
ce to deal with large plus complex datasets.
Implementing ML elements to provide regression related trends and analysis to potentially test devices on athletes
to capture data and build relevant datasets.
Data Analyst Consultant | UCF College of Engineering and Computer Science | Orlando, FL                  Aug 2020
– Present
Relevant Coursework                    Database Systems, Data Preparation, Network Science, Statistical Analysis, Int
eractive Data Visualization, Machine Learning, Data Mining, Advanced Text Mining, Parallel and Cloud Computing, Dat
a Structures and Algorithms, Business Intelligence and its Applications
Introduction to Data Analytics for Business by University of Colorado Boulder | Coursera
Built a binary classification model using Machine Learning to predict Diabetes in the United States using Python.
```

Figure 4: Summarized resume with a ratio of 10%.

As observed from the figures above, the entire text document has been summarized starting from the first line of the resume till the last line and has only taken significant semantics summarizing to 10% from the original document. This summarized document would be used to calculate the Cosine similarity and Jaccard distance in the upcoming section to predict accuracy.

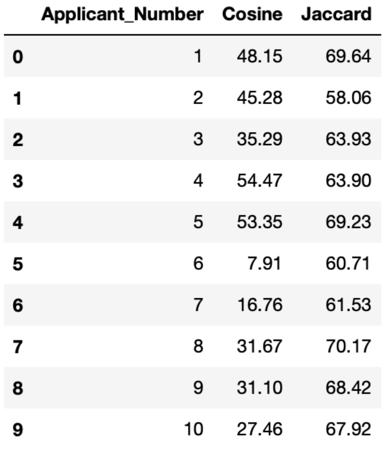
3.4 Similarity Scores using Summarized Text

The paper has used two measures of similarity and dissimilarity. The cosine similarity has been explained above as per the formula (1). The Jaccard distance is a complement of Jaccard index measures how dissimilar two datasets are by giving a percentage value accordingly. The following is the formula used to compute Jaccard index and distance:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{(|A|)+(|B|)-(|A \cap B|)} \quad (2)$$

Jaccard Distance = 1- J (A, B) (3)

Considering the formulae above, python calculated the Cosine and Jaccard similarity measures for a sample of 10 summarized resumes and following were the results.



	Applicant_Number	Cosine	Jaccard
0	1	48.15	69.64
1	2	45.28	58.06
2	3	35.29	63.93
3	4	54.47	63.90
4	5	53.35	69.23
5	6	7.91	60.71
6	7	16.76	61.53
7	8	31.67	70.17
8	9	31.10	68.42
9	10	27.46	67.92

Figure 5: Cosine and Jaccard similarity scores for 10 candidates.

It can be interpretated that Candidate 6 has the least Cosine similarity measure out of the 10 candidates, that is, 7.91% and is most likely to be rejected by an employer. However, candidate 4 has the highest Cosine similarity measure, that is, 54.47% and is likely to be accepted for the interview procedure.

Similarity, the inverse can be applied to Jaccard distance since it calculates the percentage of dissimilarity. Keeping this in mind, candidate 7 has the highest dissimilarity measure, that is, 70.17% which means that certain candidate's resume does not align well with the job requirements of the employer and is likely to be not considered for the same. It can also be seen that candidate 2 has the least dissimilarity measure, that is, 58.06% out of all the 10 candidates and is more likely to be considered for the job role.

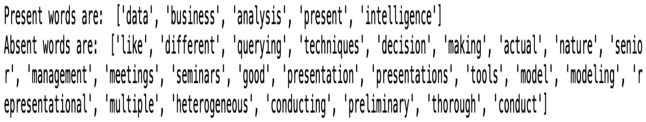
The Cosine and Jaccard similarity measures can be useful to any employer and candidate to see where the candidate is lacking by skills which will be helpful for the candidate to improve upon and likewise for the employer to choose the right talent required for the company. To use the exact skills required and missing in the resume the upcoming section sheds some light on it.

3.5 Missing Keywords

Once the candidate has an idea about how probable their resume is to be accepted or rejected for the next step in the application process, it is quite beneficial to see what further skills they need to improve upon or might have missed out on by not including in their resumes.

The same reasoning can be applied at the employers' end to analyze why they should choose a candidate or not based on the skills missing or present from their resume in order to consider them for the next interview stage.

The following is a sample output observed from a summarized resume based on an entry-level data analyst job description found on Indeed.com.



```
Present words are:  ['data', 'business', 'analysis', 'present', 'intelligence']
Absent words are:  ['like', 'different', 'querying', 'techniques', 'decision', 'making', 'actual', 'nature', 'senio
r', 'management', 'meetings', 'seminars', 'good', 'presentation', 'presentations', 'tools', 'model', 'modeling', 'r
epresentational', 'multiple', 'heterogeneous', 'conducting', 'preliminary', 'thorough', 'conduct']
```

Figure 6: Present and absent words from a resume vs job description with a ratio of 0.7 of total keywords.

It can be detected from the above figure that the labels for Present and absent words are detected from the resume by keeping a ratio of 0.7 out of the total keywords present. The ratio of 0.7 was kept for the keywords of summarized resume and original job description.

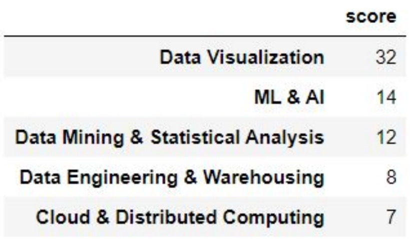
For example, as an entry level data analyst, it is highly likely that the keyword 'data' is present in the resume which matches with the keyword from the job description and is therefore under the present label. However, one can notice that 'techniques' a required keyword in the job description does not match with any of the keywords present in the summarized resume and hence comes under the absent label. The same logic can be applied for all the keywords which are present and absent by this method.

To sum up, up till this point a candidate and an employer has an idea of how much percentage of skills are present or absent but using missing keywords method, one can clearly notice on which skills the candidate should focus on or expand their knowledge about since these keywords are a prerequisite to be preferred for a job as an entry-level data analyst.

3.6 Skill-wise Classification

Text classification is a process of classifying/categorizing textual data into groups. After categorizing the data into groups and using techniques of NLP for text classifiers, it than assigns pre-defined tags/score to each of the group based on the data and then analyzes it. It has been proven effective for unstructured data to create groups/categories. It has widely been used in Sentiment Analysis, Language Detection, Topic Detection and in this paper to apply on unstructured resume and to create dictionaries from a well-defined job description based on various skills required by an entry-level data analyst / data scientist as well as assign scores to each skill based on the requirement.

The following figure represents how this paper has implemented five dictionaries on a sample of 6 resumes and has assigned a score on each one of them based on relevance.



	score
Data Visualization	32
ML & AI	14
Data Mining & Statistical Analysis	12
Data Engineering & Warehousing	8
Cloud & Distributed Computing	7

Figure 7: Skill-wise dictionaries for a data analyst along with scores for each dictionary.

It can be observed that as an entry level data analyst, the job description extracted from indeed.com requires the highest weightage of data visualization skills and hence been marked the highest score of 32. This is quite valid because given the position, the candidate should have strong foundation of data cleaning, pre-processing and basic exploratory data analysis. The least score is marked for Cloud and Distributed Computing which is at 7. This could be valid because it is a very crucial and interesting area of expertise, but one has to be perfected in the other areas of data analysis before implementing the cloud and distributing systems. The rest of the scores can be justified in a similar way for the respective skills.

Once the scores had been assigned, a resume can be mapped upon based on the dictionary and the following pie chart describes how an individual's resume behaves on it.

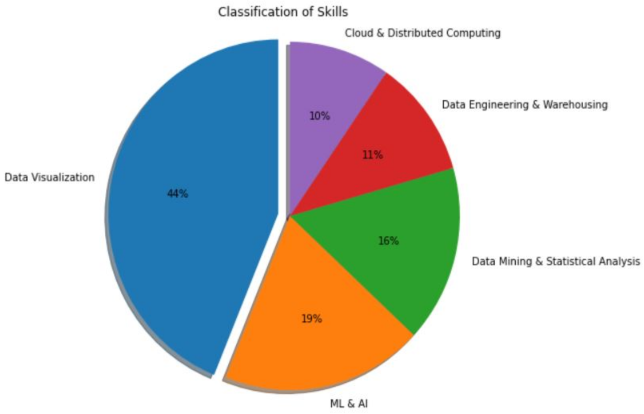


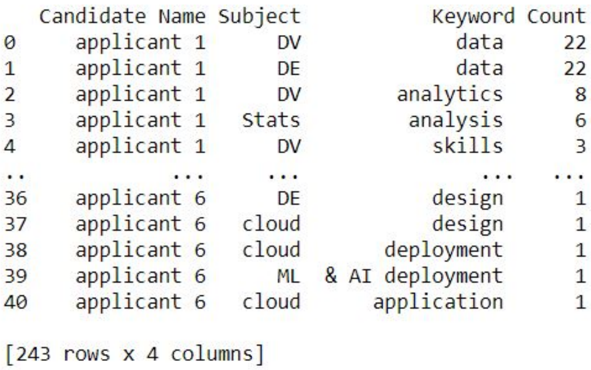
Figure 8: Pie Chart of a resume based on skill-wise dictionary

A sample resume was classified based on the dictionaries created from figure 7 and figure 8 represented how it was analyzed. It can be deducted that this candidate has the maximum data visualization skills mounting up to 44% and the lowest skills of Cloud Computing at 10% and Data Engineering at 11%. Considering the job description calls for higher visualization skills, this candidate does have a possibility to be considered for further interview rounds.

3.7 Skill-wise Comparison

Text classification has certainly helped to classify each skill based on the weightage for one sample resume. The same judgement can be applied on a larger sample of resumes since there might be many candidates applying for the same job role in this competitive job market and to understand from an employer's point of view, it would be easier to compare and contrast these skills against each other.

This paper has taken a sample of 6 resumes in total for the purpose of skill-wise classification and the following were observed.



```
   Candidate Name Subject          Keyword Count
0     applicant 1      DV             data    22
1     applicant 1      DE             data    22
2     applicant 1      DV        analytics     8
3     applicant 1   Stats         analysis     6
4     applicant 1      DV           skills     3
..            ...     ...              ...   ...
36    applicant 6      DE           design     1
37    applicant 6   cloud           design     1
38    applicant 6   cloud       deployment     1
39    applicant 6      ML  & AI deployment     1
40    applicant 6   cloud      application     1

[243 rows x 4 columns]
```

Figure 9: Keywords and counts for 6 resumes based on text classification.

Figure 9 has suggested the keyword and counts of candidates ranging from 1 to 6 based of skill-wise dictionaries. For example: candidate 1 has the keyword 'data' count of 22 for the dictionary data visualization and data engineering. Applicant 6 has the keyword 'design' and its count associated with 1 for the data engineering dictionary.

Since there exists about 243 rows and comparison by just reading the values might get confusing, for the purpose of this problem, the paper has visualized it in terms of a bar chart. Figure 10 visualizes the bar chart of the said 6 sample candidates' resumes and has classified the skills based on each of these resumes for the 5 dictionaries created from figure 7.

One can observe that candidate 1 has the highest data visualization score which is 24 out of all the candidates and candidate 4 is close second with 22. An employer who needs a future employee with high visualization skills is likely go for candidate 1. Similarly, candidate 3 has the lowest cloud computing score with 1 whereas candidate 1 has higher out of all which is a score of 7. The same can be applied for all of the skills based on which skill the employer requires the most for the candidate to be perfected in.

Figure 10 also suggests that an employer is more likely to choose candidate 1 since they triumph over all the other candidates in mostly all of the skills. Whereas candidate 3 has a lower chance of getting selected since they lack in almost all the skills as compared to the rest of the 5 candidates.

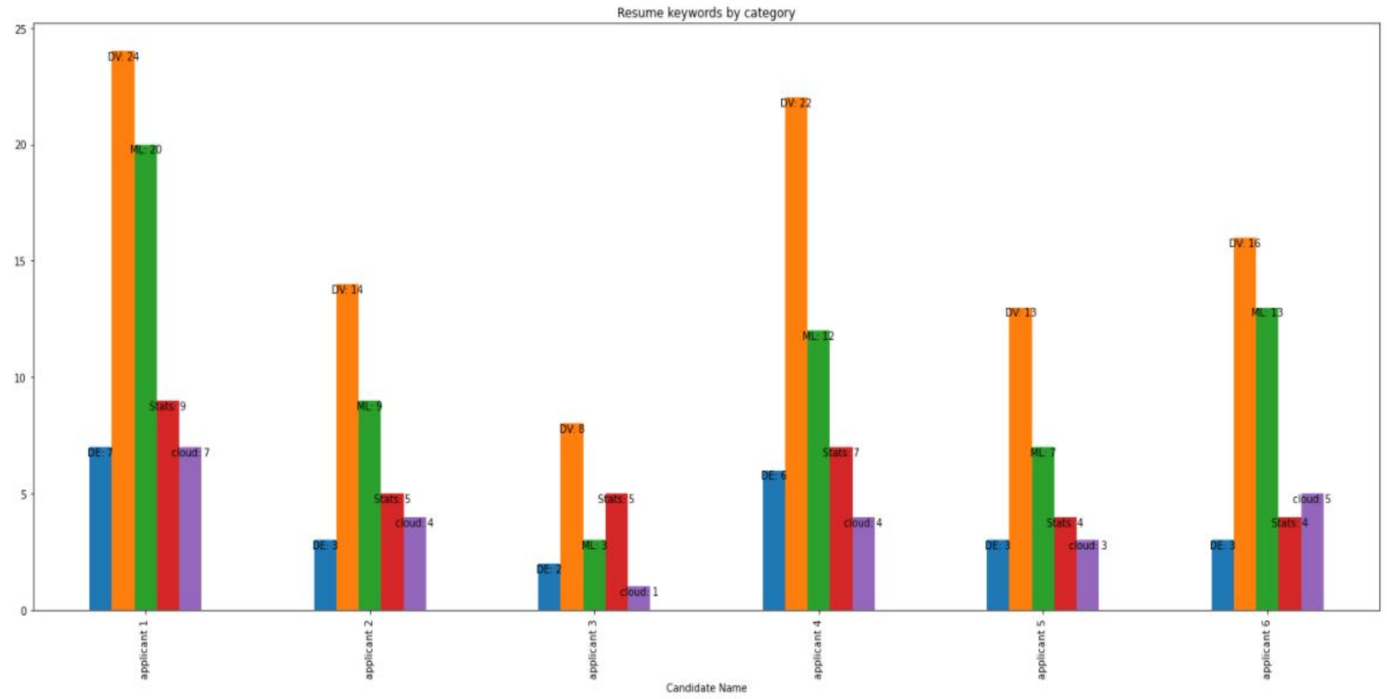


Figure 10: Bar chart for skill-wise comparison

4. CONCLUSION

The methods implemented in this project range from pattern matching, text summarization, text classification, skill-wise comparison, missing keywords and spell checker which validifies with similarity score of Cosine and dissimilarity score of Jaccard.

The paper aimed to increase the predicting power of getting rejected or accepted into a job more accurately in the future for the candidates. This will be helpful during a hiring freeze, recession, for new graduates and unpredictable pandemics. Eventually leading to low manpower and physical labor by the Human Resource departs and businesses and making everyone's jobs easier and more precise than manual selection of resumes.

It can be concluded that the paper has suggested effective methods of resume analysis which is beneficial for an applicant to analyze how well they fit in with the job description, their chances of getting accepted or rejected for a job role and the skills they can improve upon for the same. In addition, it is advantageous for an employer to choose the right candidate in terms of their necessity who will help strength their respective company in the long run.

5. LIMITATIONS

The dictionary creation tends to be biased based on the creator of the program's input. Although an employer can change the dictionary values but would require an underlying knowledge of programming to do so.

The lack of a pre-defined dataset or corpus to apply these algorithms on a much larger sample of data.

6. FUTURE WORK

More focus on missing value keywords by providing exact percentage of missingness and similarities of the job description and resumes.

The paper was focused on Data Analytics field but can be applied to any other industry roles. Extensions may include, software engineering, product management, business development, sales, etc.

Resumes can also be check for active action verbs since these verbs are more likely to occur in resume in addition to phrase matcher.

REFERENCES

[1] Shiqiang Guo, 2015, Resumatcher: A Personlized Resume Matching System. https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/154963/GUO-THESIS-2015.pdf?sequence=1

[2] Juneja Zubeda, Momin Shaheen, Gunduka Godavari, Sayed Naseem, 2016, Resume Ranking using NLP and Machine Learning. https://core.ac.uk/download/pdf/55305289.pdf

[3] CountVectorizer in Python. Educative: Interactive Course for Software Developers, 2020. https://www.educative.io/edpresso/countvectorizer-in-python.

[4] Cosine Similarity – Understanding the math and how it work? (with python). ML+, 2020. https://www.machinelearningplus.com/nlp/cosine-similarity/.

[5] Baban Deep Singh, 2020, Summarizing and Matching. In *Medium*. https://towardsdatascience.com/resume-summarizing-and-matching-165840cf9f75

[6] Venkat Raman, 2020. How I used NLP (Spacy) to screen Data Science Resumes. In *Medium*. https://towardsdatascience.com/do-the-keywords-in-your-resume-aptly-represent-what-type-of-data-scientist-you-are-591314105ba0d.

[7] Roberto Salazar, 2020. Resume Screening with Python. In *Medium.* https://towardsdatascience.com/resume-screening-with-python-1dea360be49b.