RESUME SCREENING USING MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING

S. Vairachilai, Sooraj S

School of Computing Science and Engineering, VIT Bhopal University, Madhya Pradesh – 466114

vairachilai2676@gmail.com, sooraj12sreekumar@gmail.com

Corresponding Author: S. Vairachilai, School of Computing Science and Engineering, VIT Bhopal University Bhopal-Indore Highway, Kothrikalan, Madhya Pradesh – 466114.

Abstract

The influx of applications during recruitment overwhelms companies, making manual resume screening a cumbersome task.Traditional approaches for screening resumes require manual scanning of every resume which is a time-consuming as well as an ineffective approach. This adds subjectivity and even bias into the selection process in addition to creating backlogs and delays. As a result, the sheer volume of resumes may cause qualified persons to be overlooked, making it more difficult to determine who would be the best fit for the position.This research investigates the potential of machine learning as a solution to address these limitations by developing a model to automate resume screening and enhance efficiency in identifying qualified candidates.Resume Screening is the process of applying algorithms to analyse and clarify resumes based on a particular criteria. To ease the model training, The resumes collected were preprocessed which involves methods like text normalisation, TD-IDF(Term Frequency - Inverse Document Frequency) in order to remove irrelevant information and focus only keywords which are relevant and which are related to the skills required for the job application. After the model training process is completed, the model will be able to categorise the resumes based on the job application and the skills required for it. In this research , the resume has been trained with several traditional and ensemble algorithms like Random Forest Classifier , Support Vector Machine, Naives Bayes algorithm , Logistic Regression and K Nearest Neighbours. Among the models that have been trained with the data , Support Vector Machine gives the best result with an accuracy level of 99.48%. This shows that Support Vector Machine is able to perfectly classify the resumes based on their skills and categories them according to the job application. This research can help companies to find viable applicants and hire candidates in a more effective way.

1. Introduction:

Resume screening is the first step in the busy recruitment process; it filters through a large pool of candidates to identify the most qualified for a given position. Imagine a room filled to capacity with resumes. In the past, hiring managers and recruiters would manually evaluate each application in detail to address this issue. This required carefully reviewing each résumé and contrasting the applicant's training, experience, and credentials with the requirements listed in the job posting. This comprehensive examination sought to determine whether the candidate has the skills required to succeed in the position.Several challenges faced in traditional approach:

Subjectivity and Bias: People's opinions are subjective by nature. Unconscious prejudices influenced by an applicant's name, school, or prior employers may affect selections, sometimes resulting in the rejection of deserving applicants who don't "fit the mould."

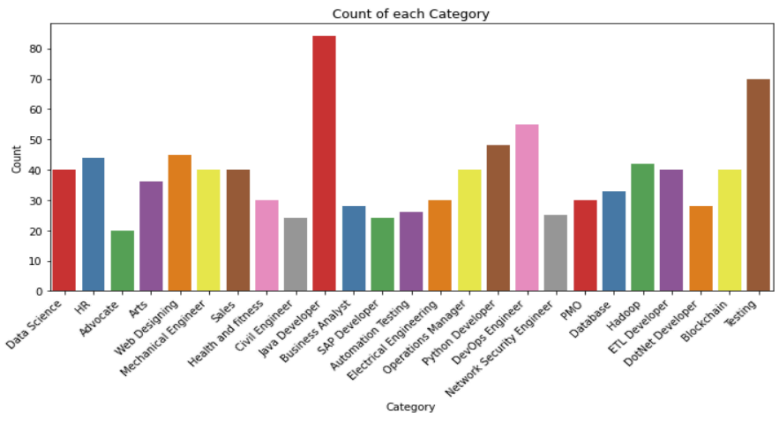
Time Restrictions: It can take a lot of time to go over every resume when there are a lot of applications. This may result in hastily completed evaluations that overlook important information about a candidate's strong points.

Some candidates use "keyword stuffing," or stuffing their applications with unrelated keywords from the job description, to get beyond screening algorithms. Though it doesn't ensure a strong fit for the position, this can fool the system.

But machine learning models are a potent ally in the era of automation. The manual keyword-based method is replaced by these intelligent resume assessors, which are complex algorithms. They probe farther, fortified by their capacity to decipher the meaning and context of a CV. This gives them the ability to determine the depth of a candidate's knowledge and competence in addition to spotting the presence of pertinent keywords.Using enormous datasets of labelled resumes and matching job descriptions, machine learning models are trained. This enables them to discover patterns that are associated with highly qualified individuals and grasp the subtleties of various roles. The programme can extract important data from a CV by analysing its content, including abilities, experience quantifiers (such years in a role, project effect), and even the language and tone utilised. Comparing this analysis to a basic keyword match provides for a more thorough examination.Consequently, machine learning models for resume screening serve as an effective gatekeeper. Employers may quickly sort through a huge applicant pool and pick applicants that best match the required qualifications thanks to its streamlined initial recruitment phase. This opens the door for additional assessment, like interviews, guaranteeing that only the most qualified candidates move forward with the employment process.

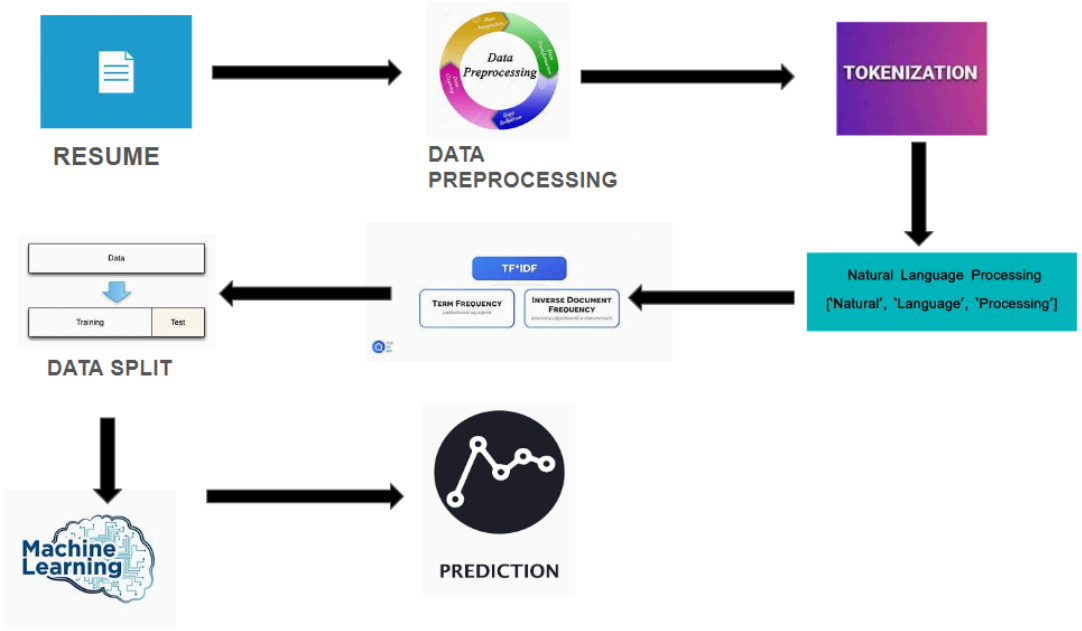
2. Dataset Description:

Two attributes make up the dataset: categories and resumes. The first contains various categories that are open for job opportunities and are classified as categorical data, while the second attribute contains the corresponding resumes that are classified as unstructured text data.



962 entries can be found in each of the dataset's two columns.The dataset contains twenty-five distinct employment categories.

3. System Architecture View:



The system adopts a machine learning approach for resume screening. First, it gathers resumes as data. This data undergoes preprocessing to ensure quality for training. Preprocessing involves techniques like tokenization (breaking text into units), lemmatization/stemming (converting words to their base form), and TF-IDF (weighting keywords based on importance). The preprocessed data is then split for training and testing a machine learning model. The system evaluates various algorithms like Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, and K Nearest Neighbors. Performance metrics like accuracy, precision, recall, F1 score, ROC curve, and AUC curve are used to determine the most effective algorithm for identifying qualified candidates from resumes.

4. Related Work:

The recruitment process in today's world has witnessed a major change with the evolution of technologies like the Internet. The following section summarises some of the literary work performed in this domain of e recruitment systems [6]. Traditional resume screening methods face several limitations. First, subjectivity and bias can creep in during human review, potentially excluding qualified candidates. Second, manually sifting through a high volume of resumes is inefficient, hindering a comprehensive evaluation.In 2019, "Automated Resume Screening System Using Machine Learning and Natural Language Processing" by Shweta Agrawal and Sumit Gupta was published in the International Journal of Innovative Technology and Exploratory Engineering. The study describes a system that uses machine learning and NLP to scan resumes and rate them based on how closely they fit the job description[2].

Machine learning is a field where we train a model with a dataset to predict the desired output when given new data. Screening the resumes is mostly done using Natural Language Processing (NLP), Natural language refers to the way we humans communicate with each other. NLP is concerned with giving computers the ability to understand the text and spoken words in much the same way human beings can[7]. With NLP, resume screening can be revolutionised by turning unstructured text into data that can be used. It may extract essential knowledge and experience, assess language for cultural fit, and even use sentiment analysis to assess soft skills, all of which contribute to a shorter and more informative shortlisting process.

Advanced NLP techniques are employed to parse uploaded resumes, extracting key information, including skills and experience, to facilitate a more accurate categorization of job seekers. Machine learning models then take centre stage to categorise candidates with exceptional precision, ensuring they are appropriately designated[4]. A machine learning model which takes the student resume as input and extracts the details like skills, certifications from it. for the extra details about the student, it also takes GitHub and LinkedIn profile links where it can extract the student contribution in various fields. The student also has to provide which job role he/she is applying for. The model is trained using a job description and skill set dataset. So ,When the resume is inputted by the student it can tell which job role is suitable for you or how your resume is relevant to the given job description[7].

Machine learning algorithms become powerful talent scouts in resume screening. Preprocessing techniques like tokenization, stemming, and lemmatization break down resumes into usable data. These techniques are then followed by methods like TF-IDF, which identifies the most relevant keywords and skills within the resumes. Finally, machine learning algorithms leverage this extracted information to categorise applicants based on their experience and fit for specific roles, streamlining the initial shortlisting process into a more efficient and data-driven system.

5. Methodology:

This project aims to explore methods for resume screening, focusing on training traditional machine learning models and ensemble algorithms to accurately categorise resumes based on a predefined criteria .

5.1 Information extraction:

The first step involves collecting information needed from the resumes to train the machine learning model. The information present in the resumes are in an unstructured format where most of the data and information available are irrelevant for the recruiters. So the aim is to collect all the relevant keywords such as skills required for the job which can be achieved using certain data preprocessing methods. There are several methods like Tokenization, stemmer, lemmatizer and TD-IDF which are used in order to extract the important keywords from the text to train the ML model.

5.2 Tokenization:

During data preprocessing, tokenization serves as the initial step, breaking down resumes into individual units of meaning. This could involve separating words, punctuation marks, or even splitting skills and experience entries containing multiple keywords (e.g., "machine learning engineer"). This process lays the foundation for further analysis by transforming unstructured textual data into a format suitable for machine learning models or statistical techniques.

5.3 Stemmer:

In data preprocessing, stemming acts as a text normalisation technique applied after tokenization. It reduces words to their base form, focusing on their core meaning. Imagine transforming "running," "runs," and "ran" into the root word "run." This helps capture synonyms and variations efficiently. While stemming might lead to some loss of information, it improves consistency and allows machine learning models to better generalise patterns within the resume text.

5.4 Lemmatization:

Lemmatization, employed in data preprocessing after tokenization, delves deeper than stemming. It strives to convert words to their dictionary or base form, considering their grammatical context. Unlike stemming, which might create unrecognisable root words, lemmatization aims for grammatically correct representations. For instance, "running" would become "run" (verb), while "runners" would be transformed to "runner" (noun). This nuanced approach enhances the accuracy of further analysis, particularly when dealing with resumes rich in verb conjugations and noun variations.

5.5 TD-IDF(TERM FREQUENCY - INVERSE DOCUMENT FREQUENCY):

In data preprocessing, TF-IDF (Term Frequency-Inverse Document Frequency) tackles the challenge of weighting keywords within resumes. It considers two factors: how frequently a term appears within a specific resume (Term Frequency) and how uncommon that term is across the entire resume dataset (Inverse Document Frequency). This combined score highlights terms that are important for a particular resume while downplaying overly common words that might appear frequently across all resumes. This weighting scheme helps machine learning models or statistical analysis prioritise relevant keywords, leading to a more accurate understanding of a candidate's skills and experience.

6. Algorithms Used:

6.1 Random Forest Classifier:

The Random Forest classifier, a robust ensemble learning technique, shines in classification tasks like resume screening. Imagine a vast forest – during training, the model builds a multitude of individual decision trees, each one analysing a random subset of skills, experience, and other features extracted from resumes. When a new resume arrives, it journeys through this entire "forest." Each tree makes its own prediction (qualified/unqualified) based on the learned decision rules. Finally, the model harnesses the collective wisdom – the most frequent classification voted on by the trees – as the final outcome for the new resume. This ensemble approach offers a significant advantage: it reduces the risk of overfitting to the training data. By incorporating the insights from multiple trees, the Random Forest classifier enhances the overall accuracy, robustness, and generalizability of the resume screening model, making it a powerful tool for efficient and effective candidate selection.

6.2 Support Vector Machine:

In the realm of machine learning classification, Support Vector Machines (SVMs) establish themselves as a powerful tool, particularly adept at tasks like resume screening. Unlike decision tree methods, SVMs don't build trees. Instead, they strategically manoeuvre in a high-dimensional space, seeking the optimal hyperplane – a dividing line – that separates the data points (resumes) most effectively. This hyperplane maximises the distance between the qualified and unqualified candidate categories. During prediction, a new resume is projected onto this feature space. Its classification (qualified/unqualified) hinges on which side of the hyperplane it falls on. SVMs excel in handling high-dimensional data and complex relationships between features, making them a valuable asset for accurate and efficient resume screening. Their ability to identify the most significant dividing line between qualified and unqualified candidates makes them a strong contender in the world of automated resume evaluation.

6.3 Naive Bayes Algorithm:

Within the machine learning classification landscape, the Naive Bayes classifier stands out for its efficiency and interpretability, making it well-suited for initial resume screening. Unlike more intricate models, Naive Bayes operates under a key assumption: the independence of features (skills, experience) within a resume. This simplifies the classification process. Imagine a detective gathering evidence – the model calculates the probability of a resume belonging to a specific category (qualified/unqualified) based on the individual probabilities of each feature, like relevant skills or keywords. By multiplying these probabilities, Naive Bayes arrives at a final classification. While based on a simplifying assumption, this probabilistic approach proves surprisingly effective, making it a valuable tool for initial resume screening. Its efficiency allows for rapid analysis of large candidate pools, and its interpretability provides insights into which resume features most influenced the classification.

6.4 Logistic Regression:

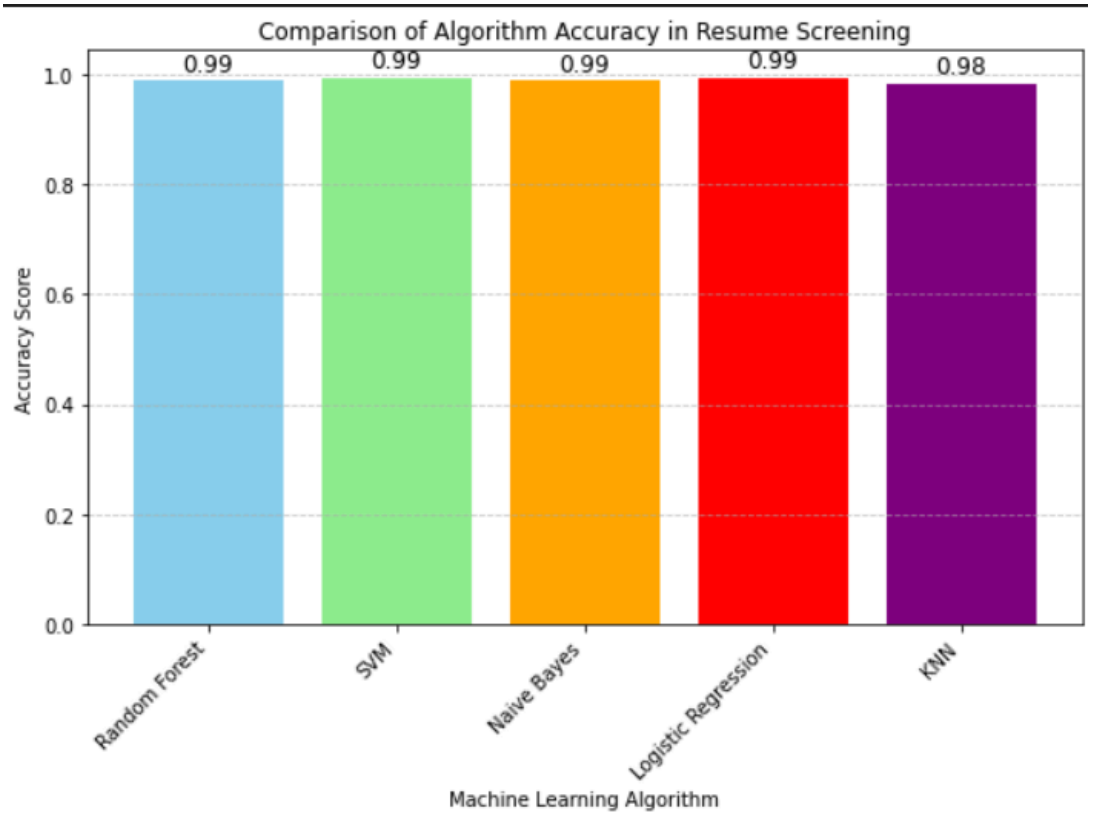
In the machine learning realm, Logistic Regression emerges as a versatile and interpretable classifier, proving its worth in tasks like resume screening. Unlike tree-based models, it doesn't build complex decision structures. Instead, Logistic Regression constructs a mathematical model that estimates the probability of a resume belonging to a specific class (qualified/unqualified) based on the presence and strength of various features (skills, experience). This model acts like a sophisticated weighing scale, assigning higher weights to features that significantly impact the likelihood of a candidate being qualified. During prediction, a new resume is fed into the model, and the calculated probability determines its classification (above a certain threshold = qualified, below = unqualified). This approach offers interpretability – you can see which resume features hold the most weight in the model's decision-making – making it a valuable tool for initial screening and understanding the key factors that define a strong candidate.

6.5 K Nearest Neighbors:

Stepping into the machine learning classification arena, K-Nearest Neighbors (KNN) offers a straightforward yet effective approach, particularly useful for initial resume screening. Unlike complex models, KNN relies on the wisdom of the crowd. During training, it stores information about existing resumes in its memory. When a new resume arrives, KNN identifies the k most similar resumes (nearest neighbors) based on their features (skills, experience). The final classification (qualified/unqualified) for the new resume is determined by the majority vote among its nearest neighbors. Imagine a group discussion – KNN gathers the opinions of the most relevant resumes (neighbors) and assigns the classification that receives the most votes. This simple approach proves surprisingly effective for resume screening, especially when the data is well-structured and the relationships between features are clear. However, KNN can struggle with high-dimensional data or noisy features, making it a strong contender for initial screening in controlled datasets.

7. Results:

This research investigated the effectiveness of various machine learning algorithms for resume screening. The project commenced by gathering a dataset of resumes, which underwent meticulous preprocessing to ensure data quality for training. This preprocessing involved techniques like tokenization, lemmatization, and TF-IDF to prepare the text for machine learning models. The preprocessed data was then divided for training and testing purposes. Subsequently, five algorithms – Random Forest, Support Vector Machine, Naive Bayes, Logistic Regression, and K-Nearest Neighbors – were employed to classify resumes as qualified or unqualified candidates. The models were evaluated based on their accuracy, achieving scores ranging from 0.984 (KNN) to 0.995 (SVM and Logistic Regression). These results demonstrate the potential of machine learning for efficient and accurate resume screening.



Because all of the algorithms utilised have greater accuracy levels—between 98% and 99%—the result demonstrates how effectively the algorithms were able to extract and identify the top resumes based on the keywords. This indicates that, in comparison to the manual method, this model can successfully select resumes that are appropriate for the firm.

8. Conclusion:

Our study looked into how machine learning might transform the way resumes are screened. We assessed the effectiveness of various machine learning algorithms, such as KNN, Random Forest Classifiers, Support Vector Machines, Naive Bayes, and Logistic Regression, by utilising a combination of pre-processing techniques like lemmatization, stemming, TF-IDF, and tokenization. This thorough investigation showed that compared to conventional keyword-based screening techniques, machine learning offers a significant improvement. Machine learning creates the foundation for a screening process that is more effective, impartial, and scalable by adjusting to the subtleties of varied resumes and doing away with human prejudice. Our findings highlight the revolutionary potential of machine learning in the future of resume screening, even if further study is necessary to further optimise model performance and address ethical problems.

9. References:

1. https://r.search.yahoo.com/_ylt=Awr1UYkD7SNmKaoulQG7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659268/RO=10/RU=https%3a%2f%2fwww.sciencedirect.com%2fscience%2farticle%2fpii%2fS187705092030750X/RK=2/RS=JwULHGw9VHrjn3qWFi3rYnc4ifY-
2. https://r.search.yahoo.com/_ylt=AwrKExRQ7SNmu1ku.li7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659345/RO=10/RU=https%3a%2f%2fwww.jetir.org%2fpapers%2fJETIR2303510.pdf/RK=2/RS=zo0qGG5az4e7sj9OeVfJeWlX..8-
3. https://r.search.yahoo.com/_ylt=Awrx_3qu7SNmlaUuNRq7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659439/RO=10/RU=https%3a%2f%2fwww.researchgate.net%2fpublication%2f361772014_RESUME_PARSER/RK=2/RS=zwhPJEc7hobdO0s2r8URQu_.1zg-
4. https://r.search.yahoo.com/_ylt=Awr1UYno7SNmhNgtMqW7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659496/RO=10/RU=https%3a%2f%2fijcrt.org%2fpapers%2fIJCRT2312507.pdf/RK=2/RS=A3Qd.Ex_zKQgrJ0jwJXY_tXyZas-
5. https://r.search.yahoo.com/_ylt=Awrx.tAr7iNm8XotmOq7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659564/RO=10/RU=https%3a%2f%2fwww.researchgate.net%2fpublication%2f221614548_PROSPECT_A_system_for_screening_candidates_for_recruitment/RK=2/RS=NkhFzdF.kcU9cJprOA9sakxG6gM-
6. https://r.search.yahoo.com/_ylt=AwrPqliL7iNmdGst1uG7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659659/RO=10/RU=https%3a%2f%2fwww.researchgate.net%2fpublication%2f347633082_AN_AUTOMATED_RESUME_SCREENING_SYSTEM_USING_NATURAL_LANGUAGE_PROCESSING_AND_SIMILARITY/RK=2/RS=JFZnQ6lX3ucvnMmCQKYYeUd6epw-
7. https://r.search.yahoo.com/_ylt=Awr1UYn27iNm4xktjUu7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659767/RO=10/RU=https%3a%2f%2fwww.semanticscholar.org%2fpaper%2fResume-Screening-using-Machine-Learning-and-NLP%253A-A-Kinge-Mandhare%2f145157f12484b32e5add1afbaecc634f4e97ff44/RK=2/RS=ky_JP3vnhAaw79NoQomrj.nKl0w-
8. https://r.search.yahoo.com/_ylt=Awrx_3pl7yNmuHMuaFK7HAx.;_ylu=Y29sbwNzZzMEcG9zAzIEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659849/RO=10/RU=https%3a%2f%2fwww.academia.edu%2f81538984%2fResume_Screening/RK=2/RS=wEtY2XvEgSr_TLe7wL5PqFWYTIM-
9. https://r.search.yahoo.com/_ylt=AwrKExTE7yNmH2QuPma7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713659972/RO=10/RU=https%3a%2f%2fieeexplore.ieee.org%2fdocument%2f9219491%2f/RK=2/RS=ySOhz6EwRz9Ml0s_wiX8Ps0jpHY-
10. https://r.search.yahoo.com/_ylt=Awr1SZuf8SNm.sluAjG7HAx.;_ylu=Y29sbwNzZzMEcG9zAzEEdnRpZAMEc2VjA3Ny/RV=2/RE=1713660447/RO=10/RU=https%3a%2f%2fieeexplore.ieee.org%2fdocument%2f10192665%2f/RK=2/RS=9c87S5yp7MPwJF1_Ss.5qHpcXi8-