A MODEL FOR AUTOMATIC RESUME SUMMARIZATION

Francisca O. OLADIPO[1]
Ajiboye A. AYOMIKUN[2]

Abstract: *Recruitment is a primary method of employing candidates for a job position and is a prime process of any organization. While this method is still the major process, it still lacks speed and accuracy of candidate choice due to the bulk and piles of resumes which are submitted for any job position. Consequently, a large number of resumes are either ignored or misplaced and only a fraction of them gets noticed. This research is aimed at developing a model to summarize the resumes by extracting important details which are essential to the recruiter and lists them in a file. The paper entails the details of the model for the summarization process and compare it with the previous methodologies. The comparison of the proposed methodology with the previous methodologies indicates the differences in the design and percentage improvement in the performance of the summarization process.*

Keywords: *model, resume, resume summarization, Natural Language Processing (NLP)*

1. Introduction

A resume is a document that an applicant submits as a first step to apply for any job opportunity. It is also an important part of the candidate application as it briefly presents the timeline of the candidate's profile. Resumes generally contain important information such as qualification, professional experience, achievements and hobbies [1]. The resume is the backbone of any job application and one important way to make leave a positive and strong impression upon recruiter [2]. However, not all resumes are well written as many applicants usually submit resumes that are very verbose and ponderous and it takes a lot of time for the recruiter to go through it and skim the important information. Since, there is no standard rule of designing and writing resumes due to which poorly crafted resumes are less appealing, hard to understand and catch the eyes of the recruiter to pick out important points from the candidature's profile. For this reason, the current search presents a model for the automatic resume summarization using the natural language processing techniques for extracting key information from a resume and listing them in a file for easy access and quick deliberation.

Automatic text summarization dates back to the almost 40 years back whose first architecture was built on IBM [13]. After which a significant work is done by Luhan which suggests to generate the summaries for the text to solve the problem

[1] Kampala International University, Uganda
[2] Federal University Lokoja, Nigeria

of understanding detailed documents. It emphasizes the importance of generating summaries of the documents as they serve as a quick tool to provide the overview of the document and thus reduces the time consumption of the reader [12]. Summarization process is a complex task which requires deep learning based natural language processing techniques [14]. In recent years, the summarization approaches using machine learning techniques have proven to improve the performance the performance of the summarization tasks [15]. But the problem with the summarization approaches is that they simply extract sentences whereas the summaries formulated by humans are more coherent and not simply a skim of sentences. This indicated the need of advanced approaches that extract passages and link them in a coherent manner. Moreover, deep phrase selection, structure building and ordering of sentences ae the problems to be addressed in the succeeding approaches [16]. The performance of text classification and data clustering is reported to be improved by a machine learning approach called as particle swarm optimization (PSO). The work of [17] has analyzed the effects of feature set on feature selection. Moreover, the significance of learned feature weights on the PSO is also analyzed. Furthermore, the learned feature weights which are produced by the PSO approach have also been used by the text summarization problem [13]. In another work, PSO approach is used to extract and classify text from the HTML web pages for [18]. Another significant work of done by the PSO approach is the document clustering which is proposed by Cui et al [19]. Web documents are reported to be classified using PSO where the terms with the highest weights are used as features for classification [20]. Apart from data clustering and classification, Swarm intelligence has also been used for the automatic text summarization where the text features are scored respective to their importance. The results of the study have shown the similarity of the summarization produced by the proposed methodology with that of MS word and suman summarization [21].

The aim of this research is to implement an automatic resume summarization model to simplify the process of recruiting that deals with resume checking. The resulting system takes resumes in different format for processing and publish a list each summarized resumes in a .txt file. In the approach deployed in this work, emails are extracted using pattern recognition using python re package; skills are mined using some pre-defined texts and are cross-checked with the predefined skills document listed in an external *csv* document.

The section 2 of the paper presents a review of the previous studies which are related to the proposed model, section 3 describes the materials and methods needed to implement the proposed model, section 4 describes the results followed by discussion and conclusion in section 5.

2. REVIEW of the RELATED WORK

Before the past century, resumes were barely in existence and jobs were awarded based on some natural qualities, such as relation, blood type (royalty or peasantry).

This is because, jobs were neither coordinated not stratified enough to warrant the use of resumes. Proper use of resume or Curriculum Vitaes never really started until after the world war, after which such use was almost so normal [3]. The summarization methods are generally categorized in three types. i.e. general linguistics, statistical and hybrid approach that is the merger of the other two categories [13].

According to [4], text summarization is the technique of making long pieces of text short with the intention to create a coherent and fluent summary having only the main points outlined in the document. Additionally, resume summarization is the shorting of a full resume or CV listing the personal information, skills and experiences. While text summarization can be from unstructured text, resume summarization can be from unstructured, semi-structured or structured text. Automatic text summarization is achieving text summarization with natural language processing in machine learning.

Some of the approaches used to extract important information from resumes include Named Entity Recognition (NER) using Natural Language Tool Kit (NLTK), where text blocks are classified into segments which are in turn chopped into words and these words are labelled using the Natural Language Tool Kit which labels by assigning tags to each word [5].

A research by [6] discussed some recent advances in Deep Learning and how the techniques is evolving and being applied in every sphere of life including document analysis. [7] described how Artificial Intelligences (AI), and the applications of Natural Language Processing (NLP) are applied in the recruitment industry and how AI is propelling significant automation across the hiring processes thereby optimizing the recruitment of quality candidates.

While a weight-based text summarization was model was proposed by [8] where sentences in a document are scored by attaching weights to the different language elements (noun, phrases, etc.); [9] applied the summarization heuristics to generate variable length summary extract from a single document. In order to evaluate the qualities of the generated summaries, the research further compared the original documents to the summaries and the experiment revealed that 65% of the documents showed less than 10% variance in scores.

Using a combination of linguistics and machine learning based approaches, [10] developed a parsing-specifications separation framework for extracting structured information from unstructured resume. The author demonstrated the process using one example and opined that it may not work for other formats.

Using three named approaches, [11] developed a resume summarization system using Python for NLP. The system was applied on 200 Data Science applicants following three steps. (i) Construction of a dictionary of all the skill sets categories, (ii) Development of an NLP algorithm to pass the whole resume and search for the words in the dictionary (iii) Count the occurrences of the target words for each candidate and aggregate.

Creating a summarization system can be very time consuming, as there are no standard algorithms for doing this and unlike other Machine Learning fields, Natural Language processing does not follow a regular pattern. Existing resume summarization systems are largely in use in industries and companies that can afford to build such systems as they work entirely based on hard-coded programming.

3. MATERIALS and METHODS

Fig 1 depicts the architecture of the proposed model for the resume summarization system. The input to this model is the domain-specific textual PDF documents of resumes that contain multifarious information of the candidate which can be fragmented into various parts i.e. personal information, qualification, experience, skills, hobbies and achievements. The features from these parts are then calculated and the performance of the model is evaluated. The platform requirements for the implementation of the proposed system are enlisted as follows.

a) A stack of resumes in the same format (preferably PDF format) in the same directory as the algorithm.
b) A working anaconda environment.
c) Installed anaconda libraries.
d) Jupyter lab
e) A working model
f) Test feed.
g) The importables: *csv, re, spacy, pandas, io, pdfminer modules– PDFResourceManager, PDFPageInterpreter, TextConverter, LAParams, PDF page – os, sys, getopt, numpy, beautifulSoup – a bs4 module, request, en-core-web-sm.*

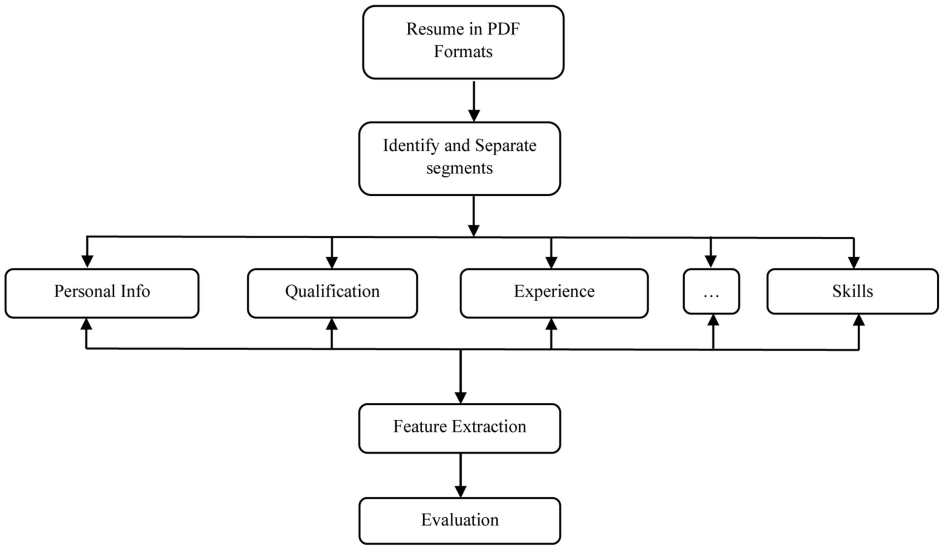

Fig 1. Design architecture

Fig 2 shows a resume sample (downloaded from the source *reddit.com)* that consists of various kind of information of the resume owner. The working of the proposed model is explained in Fig.3 which depicts the dynamic flow of the activities needed to be performed for the implementation of the proposed system model.

Jon Snow

Made at ResumeGem.com
jon.snow@housestark.raven

EXPERIENCE

Kingdom of the The North | Winterfell

Warden of the North 304 AL-Present

- Participated in strategic talks regarding the threat of the White Walkers and the Night King
- Used personal charm to strengthen alliance with Queen Danaerys Targaryen

King in the North 303 AL- 304 AL

- Led a coalition of Northerners and Free Folk to victory against House Bolton at the Battle of the Bastards
- Forged an alliance with Danaerys Targaryen, securing access to Dragonglass reserves beneath Dragonstone
- Led expedition to capture a wight, in an effort to build consensus regarding the existence of the Night King
- Restored Stark rule and began rebuilding the North after its devastation during the War of the Five Kings

The Night's Watch | Castle Black

Brother of the Night's Watch 300 AL- 302 AL

- Organized the defense of Castle Black and withstood massive assault by Mance Rayder's coalition of Free Folk
- Defeated a band of Night's Watch mutineers, avenging the death of Lord Commander Jeor Mormont
- Infiltrated the forces of Mance Rayder and gained valuable intelligence on the Wilding army

Lord Commander 302- AL- 303 Al

- Negotiated improved relations between the Free Folk and the North, Offering land in exchange for protections
- Arose from the dead after being assassinated by a nativist faction of the Night's Watch
- Rescued a band of Free Folk from Hardhome where they were besieged by The Army of the Dead

Steward to Lord Commander Jeor Mormont 298 AL-300 AL

- Chosen as Steward to Lord Commander Mormont- an indication of leadership potential
- Killed a wight that was attacking Commander Mormont. Received Longclaw, a Valyrian steel blade as reward
- Assisted in cooking, cleaning, hunting, and message delivery which was vital to Castle Black's operations.

EDUCATION

Student of Maester Luwin & Ser Rodrick Cassel | Winterfell 285 AL- 298 AL

- Trained in various arms including bow and arrow, longsword, and close combat
- Received a Highborn education studying statecraft, oration, military strategy, and Westerosi History

SKILLS & INTERESTS

Weapon: Longclaw-Valyrian Steel Blade
Pets: Ghost(Direwolf)
Language: Common Tongue
Supernatural: Can Rise From the Dead

Fig. 2. A sample resume downloaded from the source *reddit.com*

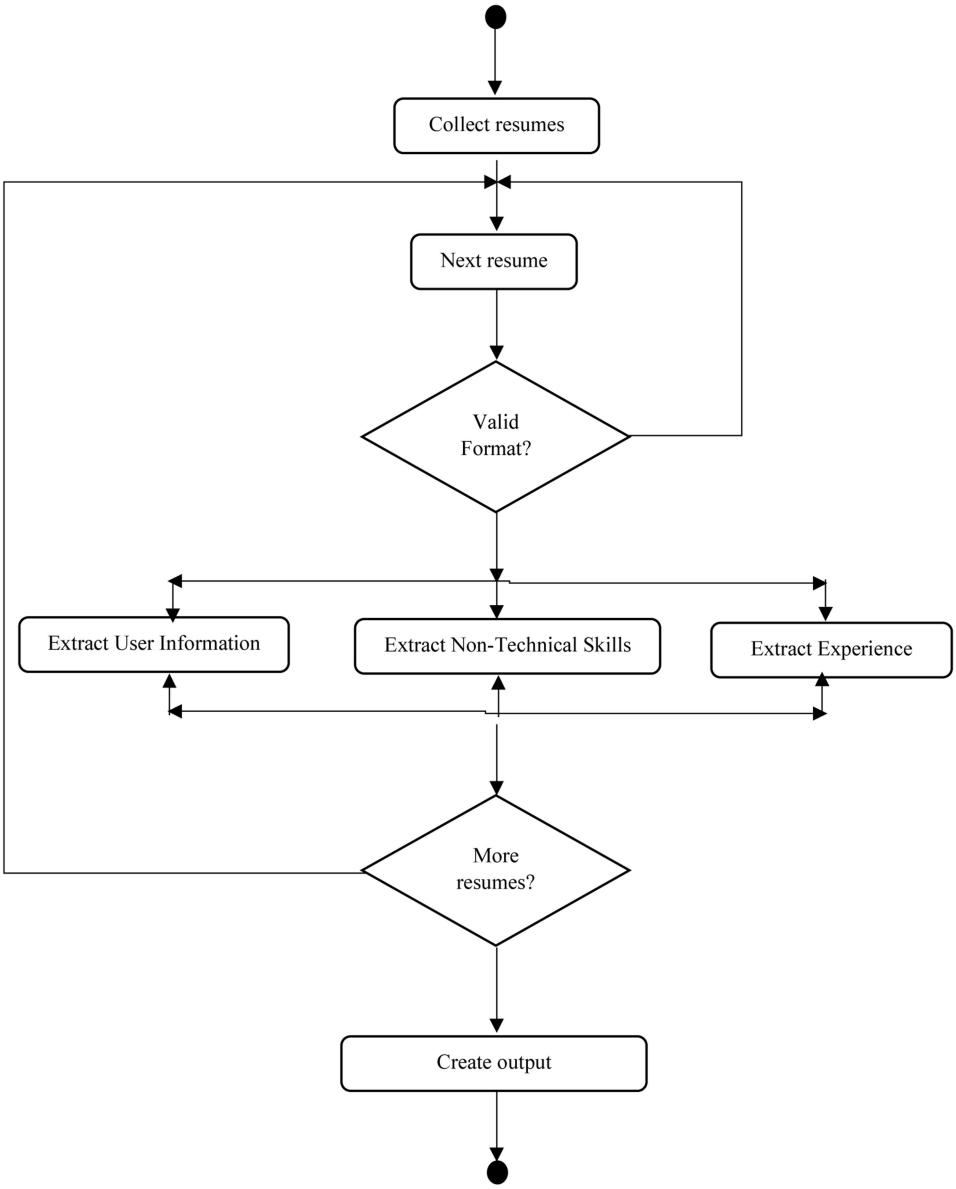

Fig. 3. Activity Flow

Fig 4 depicts the flow chart of the proposed model that describes the step by step implementation of the proposed auto resume summarization model. The Algorithmic specifications of each step are defined as follows.

01 | Start
02 | Import the importables.
03 | Select files that contain technical skills and non-technical skills

04 | Collect all resumes
05 | Pick next resume
06 | Extract Name, Phone Number and Email address.
07 | Convert the list of all predefined technical skills into an array.
08 | Check for skill in array.
09 | Save skills.
10 | Convert the list of all predefined non-technical skills into sets.
11 | Check through the resume for any non-technical skills
12 | Save the non-technical skills.
13 | Go to 05 unless resume is exhausted.
14 | Display a data frame for each and convert to csv format.
15 | End

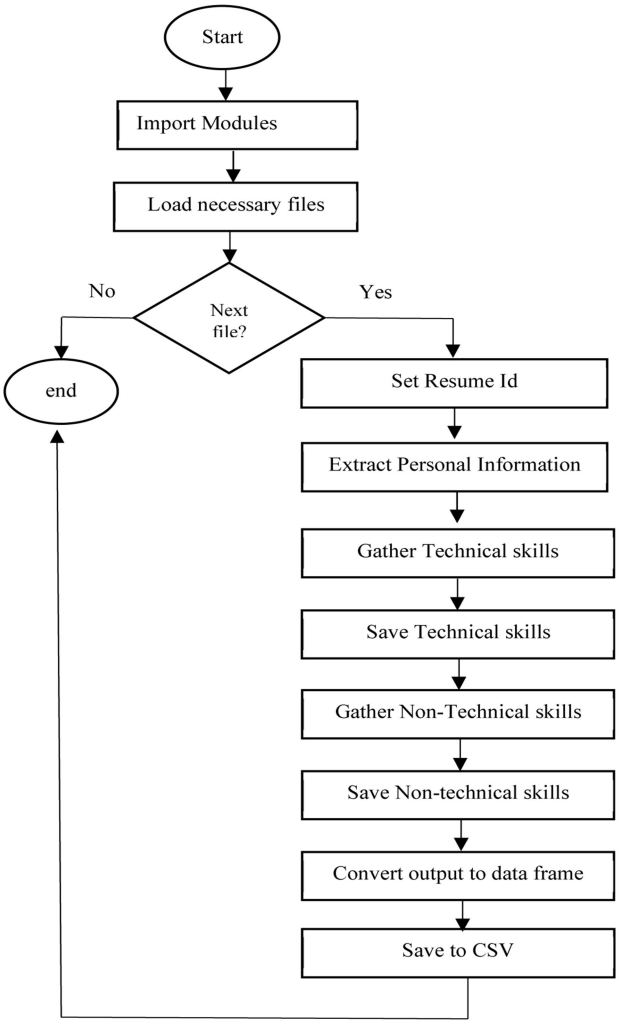

Fig 4. Flow chart of the proposed model for auto resume summarization

4. Results

The low-level specification of the summarization model consists of four integrated modules i.e. conversion, extraction general and resume list module. The detailed implementation and results of the proposed resume summarization model is described in the following subsections.

4.1 Conversion Module

Fig 5 describes the python code implementation of the conversion module that is used to convert pdf documents into plain text. The conversion module is used to convert pdf into texts. It holds to arguments which are the filename and the number of pages whose default is "None".

```
def convert(fname, pages=None):
    if not pages:
        pagenums = set()
    else:
        pagenums = set(pages)

    output = StringIO()
    manager = PDFResourceManager()
    converter = TextConverter(manager, output, laparams=LAParams())
    interpreter = PDFPageInterpreter(manager, converter)

    infile = open(fname, 'rb')
    for page in PDFPage.get_pages(infile, pagenums):
        interpreter.process_page(page)
    infile.close()
    converter.close()
    text = output.getvalue()
    output.close
    return text
```

Fig. 5. Conversion module of the resume summarization model

4.2 Extraction Modules

The extraction modules of the proposed resume summarization model are of three types i.e. name extraction module, phone number extraction module and email extraction module. Fig 6 describes the python code implementation of the name extraction module of the resume summarization model. Fig 7 describes the python code implementation of the extraction modules of the phone number and email from the resume documents. The functionality of each extraction module is described as follows.

a) **Name Extraction Module:** This module extracts the name of the resume owner.
b) **Phone Number Extraction Module:** This module extracts the phone number on the resume.

c) **Email Extraction Module:** This module extracts the email address on the resume.

```
#Function to extract names from the string using spacy
def extract_name(string):
    r1 = str(string)
    nlp = en_core_web_sm.load()
    doc = nlp(r1)
    for ent in doc.ents:
        if(ent.label_ == 'PER'):
            print(ent.text)
            break
```

Fig. 6. Name extraction module of the resume summarization model

```
#Function to extract Phone Numbers from string using regular expressions
def extract_phone_numbers(string):
    r = re.compile(r'(\d{3}[-\.\s]??\d{3}[-\.\s]??\d{4}|\(\d{3}\)\s*\d{3}[-\.\s]??\d{4}|\d{3}[-\.\s]??\d{4})')
    phone_numbers = r.findall(string)
    return [re.sub(r'\D', '', number) for number in phone_numbers]
```

Fig. 7. Phone Number and Email Extraction modules of the resume summarization model

4.3 General Module

Fig 8 describes the python code implementation of the general module of the proposed resume summarization model. This module includes all the operations that are used by the other modules.

```
def general_extraction():
    convert_non_tech_list_to_set = set(your_list[0])
    tech_list = your_list
    tech_att_list = your_listatt
    skillsatt = []


    pdfs_to_string(get_resume_list(noOfResumes))

#     for resume in resume_string_unformatted:
#         print(extract_name(resume))

    for resume in resume_string_list:
        current_skills = []
        skills_index = []
        y=extract_phone_numbers(resume)
        print('Phone Number: ', extract_phone_numbers(resume))
        y1 = []
        .
        .
        .
```

Fig. 8 General module of the resume summarization model

4.4. Resume List Module

Fig 9 describes the python code implementation of the resume list module of the resume summarization model. This module acquires all the resumes that are needed to be summarized.

```
#function to get the resume list.
def get_resume_list(noOfResuems):
    for i in range(noOfResumes):
        # all resumes have to be in pdf format.
        resumeList.append(path + 'resumes_'+ str(i) + '.pdf')
        print(resumeList)
    return resumeList
```

Fig. 9. Resume list module of the resume summarization model

5. Discussion and conclusion

A resume is an important part of the recruitment procedure of any organization as it briefly presents the timeline of the candidate's profile. Resumes generally contain important information such as qualification, professional experience, achievements and hobbies [1]. This information is important for the recruiter to document and help in making a valid decision for the candidate. For this reason, there is a need to summarize the resumes for the ease and convenience of the recruiters. Summarization of resumes is concerned with the use of natural language processing techniques for extracting key information in a resume and listing them out for easy access and quick deliberation.

Text summarization is an approach that makes the long pieces of text short with the intention to create a coherent and fluent summary having only the main points outlined in the document. Additionally, resume summarization is the shorting of a full resume or CV listing the personal information, skills and experiences. While text summarization can be from unstructured text, resume summarization can be from unstructured, semi-structured or structured text. Automatic text summarization is achieving text summarization with natural language processing in machine learning [4]. Previous studies have presented some approaches that extract the important information from resumes include Named Entity Recognition (NER) using Natural Language Tool Kit (NLTK), where text blocks are classified into segments which are in turn chopped into words and these words are labelled using the Natural Language Tool Kit that labels by assigning tags to each word [5]. Swarm intelligence is another approach that is used for analyzing the performance of text classification and data clustering using particle swarm optimization (PSO) which is also used for the text summarization problem [13]. It is also used for the extraction and classification of text from the HTML web pages [18, 20]. Web documents are reported to be classified using PSO where the terms with the highest weights are used as features for classification [20]. Apart from data clustering and classification, Swarm intelligence-based approaches i.e. PSO has also been used for the automatic text summarization [21].

Despite a lot of techniques have previously been presented for the automatic text summarization problem, however an efficient resume summarization model will require a

very large dataset of different skills in order to produce an accurate result. Developing a summarization model can be very time consuming and tedious, as there are no standard algorithms for doing this and unlike other Machine Learning approaches, Natural Language processing does not follow a regular pattern. Existing resume summarization systems are largely in use in industries and companies that can afford to build such summarization systems as they work entirely based on hard-coded programming.

This paper presents a natural language processing-based model for the auto summarization of the resumes. The input to this model is the domain-specific textual PDF documents of resumes that contain multifarious information of the candidate which can be fragmented into various parts i.e. personal information, qualification, experience, skills, hobbies and achievements. The features from these parts are then calculated and the performance of the model is evaluated. The low-level specification of the summarization model consists of four integrated modules i.e. conversion, extraction general and resume list module. Each module is implemented in Python and is used to either extract or summarize certain information of resume.

It is concluded that as the resume summarization system can be time consuming and existing machine learning approaches i.e. Natural Language processing does not follow a regular pattern due to which a generalized summarization model is not an efficient solution. It is further interpreted that the different skillset is needed in the different recruitment areas that is why the data that works for one industry might not work for another especially when the both industries specialize on different things. Hence, a generalized model for the auto summarization of resumes might not work well in all industries and more specialized or customized model is needed to work well for various industries.

REFERENCES

[1] Susan Heathfield (2018). Learn Why a Resume Is Important for an Employer.https://www.thebalancecareers.com/resume-and-why-is-it-important-1918246

[2] Mike Simpson (2019). How To Write A Killer Resume Objective. https://theinterviewguys.com/objective-for-resume/

[3] Howden, D. (2016, Februiary 4). *Die hard: the troubled history of the resume.* Retrieved from Worrkable: www.workable.com/resource/blog/recruiting/die-hard-the-troubled-history-of-the-resuem

[4] Michael J. Garbade(2018). A Quick Introduction to Text Summarization in Machine Learning. https://towardsdatascience.com/a-quick-introduction-to-text-summarization-in-machine-learning-3d27ccf18a9f

[5] Steven Bird, Ewan Klein, and Edward Loper (). Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O'Reilly. Available at http://nltk.org/book_1ed/

[6] Minar, M.R., & Naher, J. (2018). Recent Advances in Deep Learning: An Overview. ArXiv, abs/1807.08169.

[7] Oodles AI(2020). Applications of natural language processing (nlp) in recruitment. http://oodlesai.over-blog.com/2020/01/applications-of-natural-language-processing-nlp-in-recruitment.html

[8] PadmaLahari, E., Kumar, D.S. and Prasad, S. (2014) 'Automatic text summarization with statistical and linguistic features using successive thresholds', in 2014 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), IEEE, pp.1519-1524.

[9] M. K. Dalal, M. Zaveri (2011). Heuristics based automatic text summarization of unstructured text. In Proceedings of the ICWET '11 International Conference & Workshop on Emerging Trends in Technology, Mumbai, Maharashtra, India, February 25 - 26, 2011

[10] Y. H. Kulkarni (2017). Text Mining 101: Mining Information from A Resume. Downloaded June 2020 from https://www.kdnuggets.com/2017/05/text-mining-information-resume.html

[11] V. Raman (2019). How I used NLP (Spacy) to screen Data Science Resume. Towards Data Science. Accessed Septemebr 2019 from https://towardsdatascience. com/do-the-keywords-in-your-resume-aptly-represent-what-type-of-data-scientist-you-are-59134105ba0d

[12] H. P. Luhn, "The automatic creation of literature abstracts". IBM Journal of Research and Development. 2(92), 159-165, 1958.

[13] Kyoomarsi, Farshad, Hamid Khosravi, Esfandiar Eslami, and Pooya Khosravyan Dehkordy. "Optimizing machine learning approach based on fuzzy logic in text summarization." *International Journal of Hybrid Information Technology* 2, no. 2 (2009): 105-116. http://gvpress.com/journals/ IJHIT/vol 2_no2/ 10.pdf

[14] Graeme Hirst, Chrysanne DiMarco, Eduard Hovy, and Kimberley Parsons. "Authoring and Generating Health-Education Documents." In *User Modeling: Proceedings of the Sixth International Conference UM97 Chia Laguna, Sardinia, Italy June 2–5 1997*, vol. 383, p. 107. Springer, 2014.

[15] M. A. Fattah, and F. Ren, "GA, MR, FFNN, PNN and GMM based models for automatic text summarization". Computer Speech and Language. 2008. 23(1), 126-144.

[16] Inderjcet Main , the MITRE corporation 11493 Sanset Hills noad , USA , 2003 .

[17] M. S. Binwahlan, N. Salim, and L. Suanmali, "Swarm based features selection for text summarization". IJCSNS International Journal of Computer Science and Network Security. 9(1), 175-179, 2009

[18] C. Ziegler, and M. Skubacz, "Content extraction from news pages using particle swarm optimization on linguistic and structural features". IEEE/WIC/ACM International Conference on Web Intelligence. 2-5 November 2007. Silicon Valley, USA, 242-249

[19] X. Cui, T. E. Potok, and P. Palathingal, "Document clustering using particle swarm optimization". IEEE Swarm Intelligence Symposium. 8-10 June 2005. Pasadena, California, 185-191

[20] Z. Wang, Q. Zhang, and D. Zhang, "A pso-based web document classification algorithm". IEEE Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. 30 July – 1 August 2007. Qingdao, China ,659-664

[21] Binwahlan, Mohammed Salem, Naomie Salim, and Ladda Suanmali. "Swarm based text summarization." In *2009 International Association of Computer Science and Information Technology-Spring Conference*, pp. 145-150. IEEE, 2009.