Layout Aware Resume Parsing Using NLP and Rule-based Techniques

S.P Warusawithana
Faculty of Information Technology
University of Moratuwa, Sri Lanka
supunawa@gmail.com

N.N. Perera
Faculty of Information Technology
University of Moratuwa, Sri Lanka
nimnaperera98@gmail.com

R.L. Weerasinghe
Faculty of Information Technology
University of Moratuwa, Sri Lanka
reshakalakshan@gmail.com

T.M. Hindakaraldeniya
Faculty of Information Technology
University of Moratuwa, Sri Lanka
tarumadurangi97@gmail.com

G. U. Ganegoda
Faculty of Information Technology
University of Moratuwa, Sri Lanka
upekshag@uom.lk

2023 8th International Conference on Information Technology Research (ICITR) | 979-8-3503-5950-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICITR61062.2023.10382773

***Abstract* — As a result of the rapid development seen in the field of IT, there has been a surge in the number of students choosing IT field related degrees in recent years When those students try to secure job a better fob position in the field of IT, resume plays a vital role as it is often the first document a recruiter will see in the recruitment process. Therefore, this paper introduces a layout aware resume parsing system based on NLP and rule-based techniques to extract the section wise text content from the resume. This output can be used as the input for the resume content scoring model as a resume content review system to get feedback for the resume. When comparing existing methods with the proposed system, the layout of the resume would be considered in the proposed system, and it would extract content for each section. In addition to that, the proposed system would extract all the text content, but existing systems only extract the entities. In summary, this study is focused on developing a layout aware resume parsing system based on NLP and rule-based techniques to extract the section wise text content from the resume for an accurate resume review.**

Keywords — Layout aware, Resume Parser, Text extraction, NLP, Rule-based Techniques

I. Introduction

Many undergraduates in IT-related fields are searching for chances in the field of IT due to recent changes in the educational system. However, because there are few jobs in IT organizations, there is fierce competition among undergraduates for the few possibilities that are available. The resume is really important to winning this contest. Possessing a solid resume improves chances of getting the finest job. However, there is currently no adequate online alternative for receiving feedback on undergraduates' resumes. Along with having a strong resume, preparation for the interview is also helpful.

Undergraduates must go through various processes to construct a solid resume, even if it is a potent opportunity for a candidate. Reviewing the resume is the last and most difficult phase, requiring professional knowledge. When examining the past, it is clear that many undergraduates were unable to create a solid resume that matched their abilities and experience, primarily because they lacked this specialist expertise. Therefore, to avoid this gap, this paper focuses on a new approach to help the candidate to get feedback for their resume by extracting section wise content using layout aware text extraction method using NLP and rule-based techniques.

In past research work, resume parsing was done in several ways with some limitations. However, since resumes are highly structured documents, it is not enough just extracting the content but need to be aware about the layout as well. But through research it was discovered that there doesn't currently exist any system which is capable of extracting section-based content as a whole. Rather, what the existing systems do is extract only prominent phrases from the entire section [1]–[3]. But such an extraction will not be suitable for the system.

In the existing methods the system concentrates on the issue of data extraction from resumes in PDF format and suggests a hierarchical extraction method [4]. The detailed information extraction problem is approached as a sequence labeling problem in some publications, and divide a page into blocks using heuristic criteria, categorize each block using a Conditional Random Field (CRF) model [5], and then extract the detailed information. In addition to that, the usage of deep learning-based systems has given a solution for the companies for selecting the applicants for various job vacancies in consideration of their suitability. Further, few methods extract the text content from the resumes and then easily filter out the resumes for the job position by considering the included content in the resume. As the first step, after extracting the content from the resume, it performs few Natural Language Processing techniques such as tokenization, lemmatization, part of speech tagging, chunking and the named entity recognition processes for the extracted text content for further processing.

As described in the above existing methods, since there are limitations related to the accuracy of resume parsing, a new algorithm has been proposed for layout aware text extraction using NLP and rule-based techniques to increase the accuracy.

The remainder of the paper is organized as follows: The newly proposed algorithm's methodology will be explained in Section 2. Section 3 will show the evaluation of the suggested approach, comparing it to existing methods and the recently presented algorithm while highlighting its efficiency and efficacy. Following the discussion and conclusion of the paper in Sections 4 and 5, respectively.

II. Methodology

To tackle the aforementioned issue, a layout-aware extraction technique was implemented in this module. Furthermore, a multiclass classification algorithm which can predict the most appropriate section given the content, is implemented using appropriate classes such as Profile, Education, Projects and Referees to optimize the performance of layout-aware extraction.

979-8-3503-5950-3/23/$31.00 ©2023 IEEE

Additionally, a custom NER (Named Entity Recognition) component is also included which is built using the spaCy library. The NER component was instructed to identify various important entities included in resumes, such as names of the candidate, their interests, personal skills, their contact information, and their email addresses. Further, regular expressions and metadata extraction tools are also used in the system to get important entities. In addition to that, an effective algorithm is used to identify the distinct sections using spaces between bounding boxes around words.

Finally, two different JSON files representing the extracted text content from the resume parsing module will be produced as the output to a content scoring model for further processing.

A. Data Source

To gather data that would be needed to train the models, a variety of data collecting techniques had to be used. A sizable quantity of the previously chosen undergraduates'' resumes were needed in order to successfully train the proposed models. The system administrator of the industrial training platform at the Faculty of Information Technology, University of Moratuwa, Sri Lanka, provided the resume list and their contact information..

B. Dataset Preparation

After acquiring the required number of resumes the biggest challenge was annotating the resumes according to the model requirements. In order to address it, label studio software was used to annotate the content of the resume in accordance with the relevant sections. These data were then retrieved and prepared on an excel sheet along with the results.

C. Analysis and Design

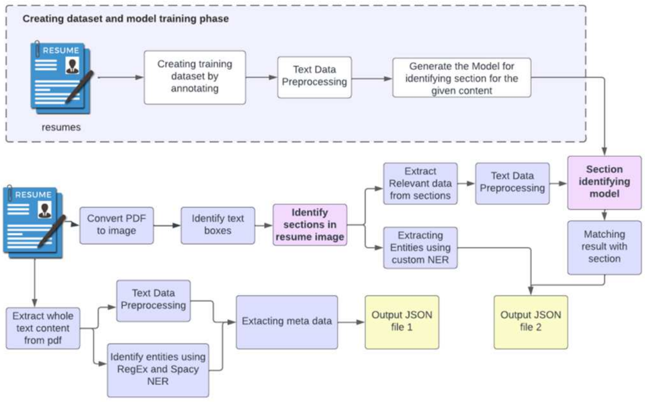
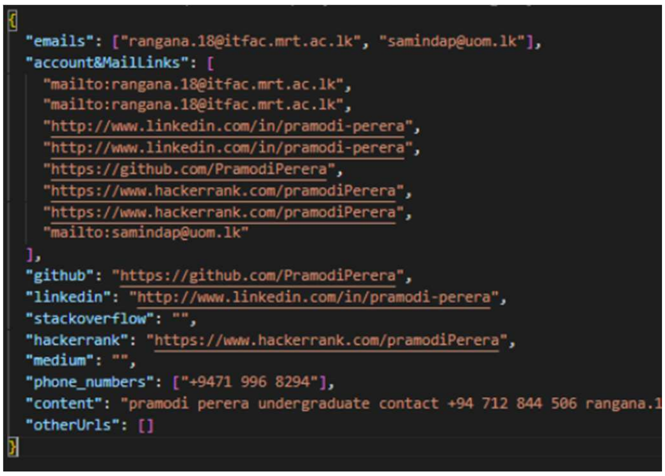

Creating dataset and model training phase
RESUME
resumes
Creating training dataset by annotating
Text Data Preprocessing
Generate the Model for identifying section for the given content
RESUME
Convert PDF to image
Identify text boxes
Identify sections in resume image
Extract Relevant data from sections
Text Data Preprocessing
Section identifying model
Extracting Entities using custom NER
Matching result with section
Extract whole text content from pdf
Text Data Preprocessing
Identify entities using RegEx and Spacy NER
Extacting meta data
Output JSON file 1
Output JSON file 2

Figure 1 High Level Diagram of the proposed system

When it comes to the development of this model, few-major approaches were considered as shown in Figure 1.

1) Extracting all text content and Entities using spaCy NER, Regular Expressions and Metadata

In the first approach, after the candidate uploads the resume to the system, all the text content from the uploaded resume was extracted using python language and the libraries. Further, several important entities such as emails, links and phone numbers also have been extracted using some rule based techniques such as regular expressions and also with spaCy NER. In order to obtain personal accounts links such as LinkedIn, Medium, and others from the resume, the system has a metadata extraction method. Figure 2 shows the output Json file which is extracted in the first approach.



```
{
  "emails": ["rangana.18@itfac.mrt.ac.lk", "samindap@uom.lk"],
  "account&MailLinks": [
    "mailto:rangana.18@itfac.mrt.ac.lk",
    "mailto:rangana.18@itfac.mrt.ac.lk",
    "http://www.linkedin.com/in/pramodi-perera",
    "http://www.linkedin.com/in/pramodi-perera",
    "https://github.com/PramodiPerera",
    "https://www.hackerrank.com/pramodiPerera",
    "https://www.hackerrank.com/pramodiPerera",
    "mailto:samindap@uom.lk"
  ],
  "github": "https://github.com/PramodiPerera",
  "linkedin": "http://www.linkedin.com/in/pramodi-perera",
  "stackoverflow": "",
  "hackerrank": "https://www.hackerrank.com/pramodiPerera",
  "medium": "",
  "phone_numbers": ["+9471 996 8294"],
  "content": "pramodi perera undergraduate contact +94 712 844 506 rangana.1
  "otherUrls": []
}
```

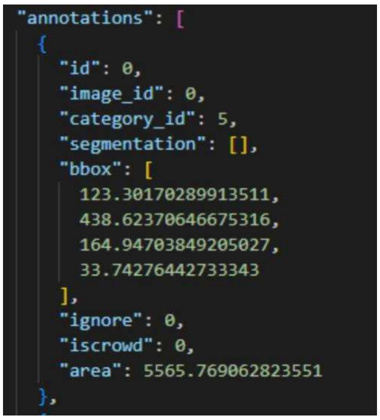
Figure 2 Extracted content as the first output.

2) Creating the dataset and Implementing the classification model

The first objective of this task is to annotate and create the required dataset for the multiclass classification model which the system needs to predict the section for the given content. Label Studio software was used for the annotation process, and it exports the final annotated data as a file called "coco" file with the relevant image files. That coco file includes a JSON file which includes all the annotation details with their positions coordination and the relevant section class label details. Figure 3 shows the format of the exported JSON file.

After exporting the annotated data from the label studio as a coco file, then a script was used to read the data from the coco file and identify the relevant sections. After identifying the relevant sections in resumes, then there should be a method to read the text content from each boundary box. Tesseract OCR was used to read the text which were included in boundary boxes.

After creating the required dataset, the main objective of the second approach is to build a multiclass classification model which can predict the section class for a given content.



```
"annotations": [
  {
    "id": 0,
    "image_id": 0,
    "category_id": 5,
    "segmentation": [],
    "bbox": [
      123.30170289913511,
      438.62370646675316,
      164.94703849205027,
      33.74276442733343
    ],
    "ignore": 0,
    "iscrowd": 0,
    "area": 5565.769062823551
  },
```

Figure 3 Position coordination of the annotations.

In the training phase the dataset which includes annotated text content and sections is used to train this model. This is a multiclass classification problem since there were several classes to predict in the current problem. As the first step, the dataset was read by using the Pandas python library and then removed all the unnecessary columns from the DataFrame. Then all the extracted text content should be prepressed before inputting that data to the model training process. Then, a pipeline was used to train the model which includes the CountVectorizer, TfidfTransformer and the Support Vector Machine [6], [7] with SGDClassifier. The SGDClassifier [8] can be used for both linear regression and support vector machine (SVM) tasks, depending on the choice of loss function and the specific parameters used. Here the system uses the SGDClassifier with Support Vector Machine algorithm.

3) Dividing section using algorithm and extracting section wise data from the resume

As the third approach, the system divides sections using algorithms and extracting section wise data from the resume.

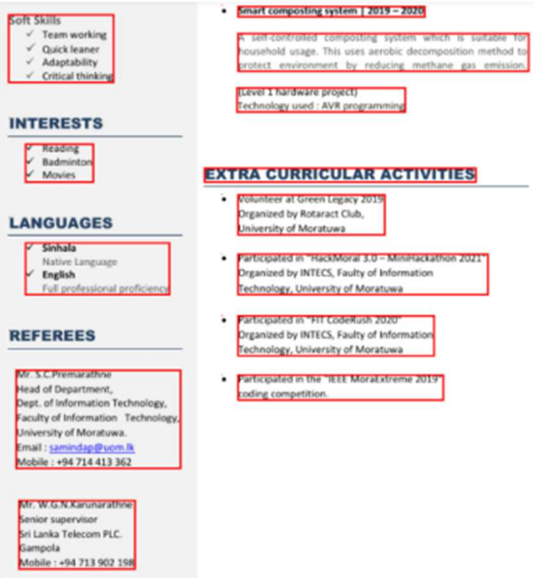
After implementing the section predicting model, in order to accurately predict the section of a given text in the resume, the system needs to detect the layouts and differentiate between paragraphs. To achieve this, an effective algorithm has been used to utilize the spacing between word boxes in the resume document. By analyzing the spaces between the word boxes, the system will identify natural breaks in the text, indicating the presence of paragraph boundaries.

To detect individual words and create bounding boxes around them in the system, the layout parser proved to be a useful tool. The layout parser enables the system to examine the spatial layout of the resume while precisely identifying and isolating each word that is used therein.

Then the system demanded the challenging work of locating distinctive paragraphs within the resume document. The system used a straightforward yet efficient approach to accomplish this goal by utilizing the bounding boxes acquired in the previous phase and the spaces between these boxes. It established a threshold value that indicates the relevance of a space as a marker of paragraph borders by carefully examining the positional information of the bounding boxes and taking the gaps between them into account.

This criterion was used to determine whether a gap between two bounding boxes was significant enough to indicate the start of one paragraph and the conclusion of another.

Through this algorithm, the system successfully delineated and differentiated various paragraphs within the resume as shown in Figure 3. This approach allowed the system to organize the resume content into coherent and distinct sections, facilitating subsequent analysis and extraction of section-specific information.



Soft Skills
- ✓ Team working
- ✓ Quick leaner
- ✓ Adaptability
- ✓ Critical thinking

INTERESTS
- ✓ Reading
- ✓ Badminton
- ✓ Movies

LANGUAGES
- ✓ Sinhala
  Native Language
- ✓ English
  Full professional proficiency

REFEREES

Mr. S.C.Premarathne
Head of Department,
Dept. of Information Technology,
Faculty of Information Technology,
University of Moratuwa.
Email : samindap@uom.lk
Mobile : +94 714 413 362

Mr. W.G.N.Karunarathne
Senior supervisor
Sri Lanka Telecom PLC.
Gampola
Mobile : +94 713 902 198

- Smart composting system | 2019 – 2020

  A self-controlled composting system which is suitable for household usage. This uses aerobic decomposition method to protect environment by reducing methane gas emission.

  (Level 1 hardware project)
  Technology used : AVR programming

EXTRA CURRICULAR ACTIVITIES
- Volunteer at Green Legacy 2019
  Organized by Rotaract Club,
  University of Moratuwa
- Participated in "HackMoral 3.0 – MiniHackathon 2021"
  Organized by INTECS, Faulty of Information
  Technology, University of Moratuwa
- Participated in "FIT CodeRush 2020"
  Organized by INTECS, Faulty of Information
  Technology, University of Moratuwa
- Participated in the "IEEE MoraExtreme 2019"
  coding competition.

Figure 4 Distinct sections identified by the algorithm.

Once the different paragraphs in the resume were recognized, the system automatically applied the trained multiclass classification model to each paragraph, essentially giving each paragraph the appropriate section. In this step, the content of each paragraph was precisely classified into categories like personal information, education, work experience, or abilities using the power of machine learning. To increase the accuracy of that, the system uses a few rule-based approaches as well.

As the summary of the methodology, in the prediction phase of the proposed system users can upload their resume as a pdf document and from the pre-trained model and the layout detection algorithms, the resume is divided into distinct sections and then the system extracts all the text content for each section. Further, extracted text content will be pre-processed and detects the relevant section from the section predicting model. Additionally, the pre-trained custom NER model will predict significant entities from the resume.

III. Evaluation

A. Evaluation of Multiclass Classification Model for Resume Section

The system used the F1-Score [9] assessment metric to evaluate the accuracy with which its multiclass classification model performed in predicting each section of a resume. The F1-Score provides an accurate evaluation of the model's classification accuracy for each section by taking into consideration both precision and recall. The system will be able to evaluate the model's performance across many areas, such as profile, education, projects, referees, awards & responsibilities, and interests by calculating the F1-Score for each section separately. This approach allowed us to gain insights into the strengths and weaknesses of the model in accurately predicting each section. The F1-Score evaluation provides a quantitative measure to evaluate the overall effectiveness of the multiclass classification model. When implementing the model, as a first step it is required to preprocess the created dataset for further usage. The dataset was cleaned by using basic NLP techniques and further by removing the section title, and additional spaces contained in the dataset. In the current implementation, a BERT transformer was utilized, hence, commonly used preprocessing steps like stemming and lemmatization were not applied to the dataset. In addition to that, the score was converted to an integer.

The system has been experimented with a few different classifiers such as Multinomial Naive Bayes, Gaussian Naive Bayes[10] and Support Vector Machine (SVM) with SGDClassifier. Then the F1-scores are calculated for each section predicted by both models as shown in Table 1. Upon analyzing the results, it became evident that the SVM with SGDClassifier consistently outperformed the Multinomial Naive Bayes classifier in terms of F1-scores for each section. The SVM with SGDClassifier exhibited higher accuracy, precision, and recall values, resulting in superior overall F1-scores across the different resume sections as shown in the Table 2.

Table 1 F1 Scores for different sections

Section	SVM (F1-Score)	NaiveBayes (F1- Score)
Profile	0.90	0.80
Education	0.91	0.82
Projects	0.92	0.90
Referees	0.97	0.98
Extra-Curricular/ Roles and Responsibilities	0.79	0.76
Awards Achievements	0.85	0.85
Personal Skills	0.88	0.73
Technology/Technical Skills	0.94	0.91
Work Experience	0.39	0.00
Interests	0.88	0.78

Table 2 Accuracy value for the different classification algorithms.

Model	Accuracy Obtained
SVM with SGDClassifier	0.8881431767337807
Multinomial NaiveBayes	0.8400447427293065

B. Cosine Similarity Analysis for Section Content Comparison

The cosine similarity metric is used to evaluate how closely the extracted relevant section to the actual section content in order to evaluate the system's accuracy and efficacy. The system determined the cosine similarity score for a set of resumes by contrasting the textual representation of the extracted section content with the actual section content. The system then calculates the average values of cosine for each section as shown in Table 3. The cosine similarity considers the direction and magnitude of the vectors representing the two texts, providing a measure of their similarity.

Table 3 Average Cosine Values for each section

Section	Average Cosine Value
Profile	0.8848024768311141
Education	0.7301637245094814
Projects	0.7836110811709401
Technical Skills	0.7353279814496185
Personal Skills	0.8700204069920932
Interests	0.740757026121876
Awards and Responsibilities	0.7716714248985228
Referees	0.8270679024187793

IV. Discussion

The main objective of this module was to find a layout-aware content extraction technique to accurately extract the section wise text content from the resumes and grab the important entities from the resumes. Through the experiments and the implementations, an effective resume parsing system has been found as the main output of this module. The problems of the existing systems were the accuracy and inability to extract section-wise data. Those systems only extract the important entities from resumes, not the whole content. The accuracy of the implemented parsing system showcased remarkable performance in accurately extracting and categorizing different sections of resumes. Through the implementation of advanced classification models and algorithms, the system demonstrated its ability to handle a diverse range of resume formats and layouts, ultimately yielding reliable results. The accuracy of the section predicting multiclass classification model which was implemented using SVM is around 0.8 which has a higher value. When considering the average cosine similarity value for each section it gives more than around 0.75 for every section. The system's exceptional accuracy in extracting section-wise information can be attributed to the successful integration of machine learning techniques, such as the multiclass classification model and custom NER.

When considering the limitations of the implemented model, the variety of resume formats makes it difficult to effectively parse and extract information. Because different people have different preferences, there are many different forms and layouts for resumes. The system may have issues when dealing with unusual or highly customized layouts, despite being built to properly handle typical resume formats. In addition to that, even when the algorithm successfully extracts section wise content, it could not have a complete knowledge about the context of that content.

As further work of this module, deep learning methods can be implemented to improve the effectiveness of our resume parsing system. Then the system will be able to learn more about the semantic and contextual information contained in resumes by investigating the use of deep learning models like recurrent neural networks (RNNs) or transformer-based models. Further, researching the application of layout-aware object detection and image processing techniques has the potential to significantly enhance the accuracy of paragraph division in the implemented resume parsing system. Then the system will be able to identify the visual layout components in resumes, including headings, subheadings, and text blocks, by utilizing advanced computer vision techniques and object detection models.

V. Conclusion

In conclusion, this study has taken on the problem of tackling an important issue in the area of layout-aware text extraction for resumes. Existing techniques have frequently ignored the crucial element of resume formatting in favor of extracting primary entities from resumes. But the suggested system, on the other hand, is a ground-breaking strategy that carefully considers the format of the resume. It makes an effort to assure the accuracy and comprehensiveness of the extracted content by using a section-wise text content extraction approach.

However, like any technological solution, it is not without its limitations. Notably, the system may encounter difficulties when processing resumes with unconventional layouts or intricate formatting. Furthermore, its performance might vary when handling resumes in languages with complex scripts or those containing extensive graphical elements.

To overcome these limitations and further improve the system's capabilities, future research avenues beckon. The incorporation of deep learning techniques, coupled with access to extensive and diverse datasets, holds immense potential for refining the accuracy and adaptability of our system. By embracing these advanced methodologies, we anticipate that subsequent iterations of our system will be better equipped to navigate the intricacies of modern resumes, thus enhancing its utility in the dynamic landscape of resume analysis.

Acknowledgment

This study is supported by the Faculty of Information Technology of University of Moratuwa under the supervision of the Department of Information Technology and Department of Interdisciplinary Studies.

References

[1] D. Chandola, A. Garg, A. Maurya, and A. Kushwaha, "Online resume parsing system using text analytics," *J. Multi Discip. Eng. Technol.*, pp. 1–5, 2015, [Online]. Available: www.jmdet.com

[2] J. Of, A. Education, O. Learning, and N. Hampshire, "See discussions, stats, and author profiles for this publication at: http://www.researchgate.net/publication/233925391," vol. 10, no. JANUARY 1993, pp. 18–23, 2015, doi: 10.13140/RG.2.2.11709.05607.

[3] C. Daryani, G. S. Chhabra, H. Patel, I. K. Chhabra, and R. Patel, "an Automated Resume Screening System Using Natural Language Processing and Similarity," vol. 2, no. July, pp. 99–103, 2020, doi: 10.26480/etit.02.2020.99.103.

[4] V. Bhatia, P. Rawat, A. Kumar, and R. R. Shah, "End-to-End Resume Parsing and Finding Candidates for a Job Description using BERT," 2019, [Online]. Available: http://arxiv.org/abs/1910.03089

[5] J. Chen, L. Gao, and Z. Tang, "Information extraction from resume documents in PDF format," *IS T Int. Symp. Electron. Imaging Sci. Technol.*, pp. 1–8, 2016, doi: 10.2352/issn.2470-1173.2016.17.drr-064.

[6] Y. Ahuja and S. Kumar Yadav, "Multiclass Classification and Support Vector Machine," *Glob. J. Comput. Sci. Technol. Interdiscip.*, vol. 12, no. 11, pp. 14–19, 2012, [Online]. Available: https://globaljournals.org/GJCST_Volume12/2-Multiclass-Classification-and.pdf

[7] J. K. Sahoo and A. Balaji, "Optimizing support vector machines for multi-class classification," *Commun. Comput. Inf. Sci.*, vol. 721, pp. 393–398, 2017, doi: 10.1007/978-981-10-5427-3_42.

[8] C. J. Varshney, A. Sharma, and D. P. Yadav, "Sentiment analysis using ensemble classification technique," *2020 IEEE Students' Conf. Eng. Syst. SCES 2020*, pp. 12–17, 2020, doi: 10.1109/SCES50439.2020.9236754.

[9] M. Grandini, E. Bagli, and G. Visani, "Metrics for Multi-Class Classification: an Overview," pp. 1–17, 2020, [Online]. Available: http://arxiv.org/abs/2008.05756

[10] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification," *2019 Int. Conf. Autom. Comput. Technol. Manag. ICACTM 2019*, pp. 593–596, 2019, doi: 10.1109/ICACTM.2019.8776800.