Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2024) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

6th International Conference on AI in Computational Linguistics

ResuméAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models

Ahmed Heakl[a], Youssef Mohamed[1], Noran Mohamed[a], Ali Sharkaway[a], Ahmed Zaky[a]

[a]*Egypt-Japan University of Science and Technology, New Borg El-Arab City, 21934, Alexandria, Egypt*

Abstract

The increasing reliance on online recruitment platforms coupled with the adoption of AI technologies has highlighted the critical need for efficient resume classification methods. However, challenges such as small datasets, lack of standardized resume templates, and privacy concerns hinder the accuracy and effectiveness of existing classification models. In this work, we address these challenges by presenting a comprehensive approach to resume classification. We curated a large-scale dataset of 13,389 resumes from diverse sources and employed Large Language Models (LLMs) such as BERT and Gemma1.1 2B for classification. Our results demonstrate significant improvements over traditional machine learning approaches, with our best model achieving a top-1 accuracy of 92% and a top-5 accuracy of 97.5%. These findings underscore the importance of dataset quality and advanced model architectures in enhancing the accuracy and robustness of resume classification systems, thus advancing the field of online recruitment practices.

Keywords: Resume Classification; Online Recruitment; Dataset Collection; Large Language Models (LLMs); Transformers

1. Introduction

In today's fast-paced and competitive job market, online recruitment has become integral to employers and job seekers. With the global online recruitment market projected to reach USD 39.76 billion in 2022 and expected to grow at a Compound Annual Growth Rate (CAGR) of 7.2% between 2021-2026, understanding the intricacies of this evolving landscape is paramount [18]. One crucial aspect of online recruitment that demands attention is resume classification [4]. As nearly 70% of companies utilize online recruiting platforms and 94% of employers plan to adopt or continue using AI for talent acquisition and HR, the volume of job applications is skyrocketing [28]. Consequently, efficiently categorizing and analyzing resumes has emerged as a critical challenge for recruiters.

Despite the growing reliance on online recruitment platforms and the adoption of AI technologies, several challenges persist in resume classification. One significant hurdle is the availability and quality of datasets for training AI algorithms. The collection of datasets from recruiting platforms poses challenges due to the sensitive nature of the data, which often contains personal information about job seekers [29]. Ensuring data privacy and compliance with regulations such as GDPR (General Data Protection Regulation) adds complexity to the dataset collection process. Additionally, the lack of standardized templates for resumes further complicates resume classification efforts. With no

universal guidelines for what constitutes a high-quality resume, recruiters often encounter a plethora of resumes varying in format, structure, and content quality [30]. The variability in resume quality poses challenges for AI algorithms, potentially leading to biases and inaccuracies in candidate selection. [31]. Addressing these challenges is crucial for advancing resume classification and improving online recruitment efficiency.

A prevalent problem in the current state-of-the-art of resume classification is the reliance on small datasets with limited samples and labels. Many researchers and practitioners in the field often work with datasets containing only a few thousand samples and a small number of labels, typically ranging from 5 to 25 categories. This scarcity of data presents significant challenges for training robust and accurate classification models. Moreover, the use of classic machine learning algorithms such as Naïve Bayes [27], Support Vector Machine (SVM) [21], Random Forest [22], K-Nearest Neighbor [23], and Logistic Regression [24], alongside TF-IDF vectorization [25] and XGB algorithms [26], exacerbates the issue. While these algorithms have proven effective in various contexts, their performance may be suboptimal when confronted with large datasets and complex classification tasks [32], leading to challenges in achieving high accuracy and generalization. As such, overcoming the limitations imposed by small datasets and exploring innovative approaches to improve model performance remains a critical area of research in the field of resume classification.

The contributions of this work represent a significant advancement in the field of resume classification, addressing key challenges and pushing the boundaries of existing methodologies. Our contributions are as follows:

- **Large-Scale Dataset Collection and Preprocessing** The curated dataset, comprising 13,389 records from diverse sources across 43 distinct resume categories, represents the largest collection for resume classification to the best of our knowledge, underpinned by approximately 400 hours of meticulous data preprocessing efforts aimed at ensuring high-quality samples and minimizing noise and inconsistencies, thereby enhancing the reliability and robustness of the model.

- **Outperforming State-of-the-Art Models** We employed cutting-edge LLMs, including Gemini and BERT, to achieve remarkable accuracy, with 91% top-1 accuracy and 97% top-5 accuracy, surpassing existing state-of-the-art models in resume classification tasks.

- **Providing High-Quality Codebase** We provide high-quality codes for scraping, preprocessing, and training, facilitating reproducibility and enabling researchers to build upon our work.

To overcome limitations in resume classification, we took a comprehensive approach to dataset collection and model development. We gathered a high-quality dataset from diverse sources, including Google, Bing, and leading resume websites. This endeavor involved meticulous data collection and filtering processes, totaling approximately 400 hours, to ensure the inclusion of diverse samples representing a wide range of resume categories. As a result, we assembled the largest dataset in the context of resume classification, comprising 13,389 records across 43 classes. Leveraging state-of-the-art transformer models and Large Language Models (LLMs) such as Gemma [20] and BERT [19], we tackled the complexities of resume classification with a focus on achieving robustness and high accuracy. By employing advanced techniques and leveraging the richness of the collected dataset, our approach transcended the limitations of traditional machine learning models, resulting in the development of a highly effective and reliable model that outperforms the shortcomings of existing state-of-the-art models.

2. Related Works

Pal et al. (2022), [1] addressed the imperative for automating the hiring process amid the transition to remote work precipitated by the epidemic. It confronts the central issue of inefficiency and manual effort inherent in resume categorization and verification, proposing automation as a substantial alleviating measure. Leveraging machine-learning algorithms, including Naïve Bayes, SVMs, and random forest, the investigation endeavors to extract skills and categorize resumes into pertinent job profile classes. Data acquisition spans various sources such as Kaggle, Glassdoor, and Indeed, yielding unstructured datasets subjected to cleaning, classification, and storage processes. A training dataset encompassing 70% of the total dataset is utilized. Findings reveal the superiority of the random forest model over Naïve Bayes and SVM, yielding an accuracy rate of 70% and exhibiting enhanced predictive capabilities.

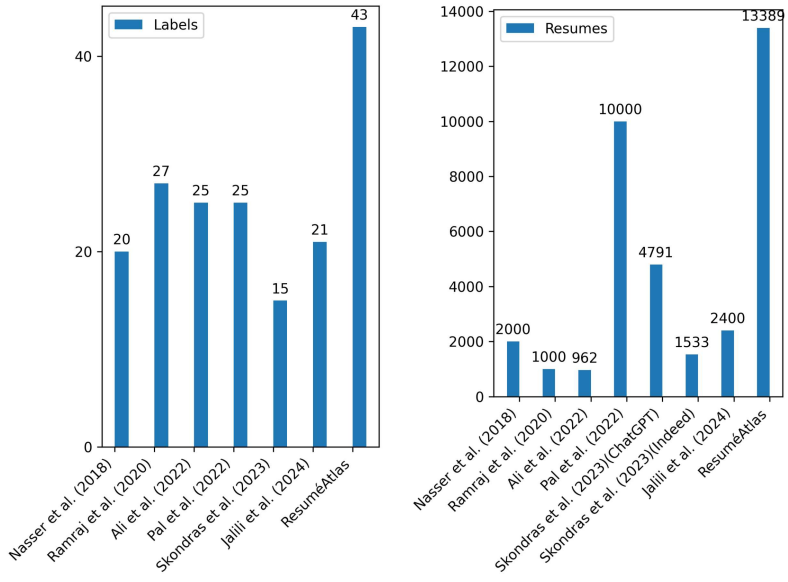
Skondras et al. (2023) [2] explored using synthetic resumes to augment training data and improve the effectiveness of resume classification algorithms, particularly in categories with sparse samples. The authors employed the OpenAI API to generate structured and unstructured resumes and resumes from the Indeed website with total records of 4791 in the ChatGPT Dataset and 1533 in the Indeed Dataset with 15 categories. These synthetic resumes were used to train two models: a transformer model (BERT) and a feedforward neural network (FFNN) incorporating Universal Sentence Encoder 4 (USE4) embeddings. The experiments showcased that the BERT mode, coupled with augmented datasets, demonstrated superior performance compared to the FFNN model. A 92% accuracy was achieved in the sixth experiment, where the BERT model was combined with the Indeed augmented dataset.

Additionally, Ali et al. (2022) [3] presented an NLP-based approach to classify resumes into job categories. The study employs a dataset of 962 labeled resumes across 25 job categories and evaluates nine ML classification models, including SVM, Naïve Bayes, and Logistic Regression. The results show that SVM classifiers, particularly the Linear Support Vector Classifier, achieve an accuracy of over 96%.

Jalili et al. (2024) [4] presents a novel method for resume classification using a Bidirectional LSTM (BiLSTM) architecture to enhance the accuracy and efficiency of candidate evaluation in the recruitment process. Their method includes text preprocessing steps, followed by utilizing BiLSTM to capture both past and future contexts of resume content. Word embedding is used to enrich the textual representations. The dataset is obtained from LiveCareer and comprises a collection of over 2400 resumes, categorized into 21 distinct job categories. The BiLSTM is good at capturing sequential dependencies achieving 72.4% in classification accuracy.

Nasser et al. (2018) [5] concentrated on the document classification domain, specifically focusing on resume categorization into distinct classes. The proposed methodology utilizes Convolutional Neural Networks (CNN) with Glove-Word Embedding for resume classification. Resumes undergo hierarchical segmentation, and a CNN model with word embedding is employed at each level for classification. The outputs from individual classifiers are amalgamated to define the overarching resume category hierarchy. The study utilizes datasets from Calpine Lab's resume collection and sample job descriptions covering both technical and non-technical domains. Eight classifiers are used for the classification task, covering binary and multi-class classifiers. The classification hierarchy spans five levels, with each level tailored to specific tasks and employing dedicated CNN models. The dataset includes multiple labels across 20 categories distributed across hierarchy levels. The CNN architecture comprises embedding, convolutional, max-pooling, dropout, and dense layers, and the system's efficacy is assessed using precision, recall, and f-score metrics. Results demonstrate promising performance, with high accuracy levels observed across different hierarchical levels. For instance, at Level 1, the training accuracy attained 99%, while the test accuracy achieved 94%. Similarly, at Level 5, the training accuracy recorded 98.7%, and the test accuracy reached 92.9%. These findings underscore the effectiveness of the CNN-based approach in accurately classifying resumes into specific categories, thereby facilitating the recruitment and selection processes of candidates.

Within the online job recruitment domain, the precise categorization of job postings and resumes holds significant significance for both job seekers and recruiters alike. Ramraj et al. (2020) [6] introduced an automated text classification system tailored to classify textual data, specifically resumes, into diverse categories utilizing advanced methodologies such as Term Frequency-Inverse Document Frequency (TF-IDF) with CNNs. The dataset comprises over 1000 resumes obtained from LinkedIn through web scraping techniques and API tools. Each resume includes attributes like job title, description, skills, location, and past experiences. Resumes are segmented into multiple labels based on job titles and descriptions, resulting in a dataset with 27 categories. Domain adaptation techniques are used due to the sensitive nature of resume data. Preprocessing ensures data quality and consistency, with resumes classified into multiple labels based on job titles and descriptions. The CNN algorithm, typically used for image classification, is adapted to extract character-level features from LinkedIn profiles. Various algorithms, including SVM, Naive Bayes (NB), and TF-IDF, are applied individually and collectively to the dataset. Outcomes reveal the CNN algorithm's superiority, manifesting in an accuracy rate of 68% and the highest F1-score of 0.65 among all evaluated models. Additionally, TF-IDF coupled with NB, SVM, and XGB algorithms also demonstrates competitive performance, yielding F1-scores ranging from 0.57 to 0.61. This study contributes valuable insights into the effectiveness of different algorithms for resume classification, offering implications for improving online job recruitment processes.



(a) Number of the labels in each dataset.

(b) Number of the collected resumes.

Fig. 1: Collected dataset samples and labels comparison with other datasets.

3. Dataset

We collected ”ResuméAtlas”, a dataset comprised of resumes sourced from Google Images, Bing Images, and LiveCareer, with textual content extracted using optical character recognition (OCR). A total of 13,389 records were obtained through automated scraping techniques, including 3,015 from Google Images, 2,722 from Bing Images, and 7,652 from LiveCareer. A comparison of the number of collected resumes and labels in this dataset to others is presented in figures 1a and 1b.

3.1. Data Collection Process

The collection process involved several stages, starting with the scraping of resume images from each source. This process was executed using separate scripts tailored for Google Images, Bing Images, and LiveCareer. The scraping phase required approximately 5 hours for Bing, about 25 hours for Google, and nearly 40 hours for LiveCareer due to the websites’ different complexities and the data volume. Following the scraping phase, the downloaded images underwent extensive filtering procedures over approximately 80 hours to ensure the quality and relevance of the data by iterating over the downloaded resumes one by one to delete the irrelevant ones. In the final stage of data preprocessing, Optical Character Recognition (OCR) algorithms were employed to extract textual content from the resume images. This process was performed separately for the Google/Bing and LiveCareer datasets, requiring approximately 95 hours and 145 hours, respectively, to process the entire dataset. Specifically, Google’s Cloud Vision service was utilized to facilitate the OCR process.

3.2. Challenges in the Dataset

The dataset presents several challenges that must be addressed during analysis. These challenges include the lack of a structured format, making it difficult to extract specific information such as name, education, and experience. Additionally, resumes may contain personal information such as names, phone numbers, and email addresses scattered throughout the text, posing privacy concerns and necessitating the implementation of data anonymization techniques. Skills or experiences may be mentioned multiple times in different sections of the resume, leading to redundancy and requiring techniques to identify and remove duplicates. Misspellings, such as ”experiance” instead of ”experience” or ”technicl” instead of ”technical”, may be present in the text, affecting the quality of the data and necessitating the

implementation of spell-checking algorithms. Headers, footers, and contact information at the top or bottom of the resume may not be relevant to the content analysis. They should be filtered out to prevent noise and improve data quality. Some resumes may contain watermarks or highlighted sections that need to be removed or accounted for during analysis to prevent bias and ensure accurate results. Special characters such as '*', '/', '&', '$', '%', '^', '~' may be present in the resume text and need to be handled appropriately during preprocessing. Resumes may also contain URLs or links to personal websites or online profiles. Finally, resumes may contain experience irrelevant to the job being applied for, which might need to be identified and filtered out during analysis to improve the accuracy of the results. These challenges highlight the importance of developing robust preprocessing and analysis techniques to ensure the quality and accuracy of the results.

3.3. Data Preprocessing

In the data preprocessing phase, we employed a series of steps to ensure the cleanliness and uniformity of the text data. We began by converting all text to lowercase to promote consistency. Next, we utilized regular expressions to remove punctuation marks and non-alphanumeric characters systematically. We also removed URLs, Twitter handles, hashtags, and special characters, and expanded contractions. Finally, we removed common stop words, such as 'and', 'the', and 'is', using the NLTK [12] stop words corpus, thereby enhancing the quality of the tokenized text.

4. Methodology

Resume classification is a fundamental task in natural language processing, where the goal is to assign a relevant class label to a given text input $c_{pred} = \arg\max_{c_i \in C} P(c_i|w_0, w_1, \ldots, w_n, \theta)$ where c_{pred} represents the predicted class label, C denotes the set of possible class labels, and $P(c_i|w_0, w_1, \ldots, w_n)$ represents the probability of the i^{th} class label given the input text, which is composed of words $w_0, w_1, \ldots, w_n$. Here θ represents the model weights which is optimized through training.

In this study, we leveraged the sequential nature of the resume classification problem to employ large language models (LLMs) as our classification approach. Specifically, we explored two prominent variants of LLMs: Bidirectional Encoder Representations from Transformers (BERT) and Gemma1.1 2B [7], [8].

For the BERT-based approach, we utilized an encoder-based architecture that receives a sequence of words as input, which are then converted into tokens and embedded. Positional embeddings were added to the textual embeddings to preserve sequential information. The model employs multi-head self-attention blocks to extract meaningful representations from the input sentences. The output of these stacked attention blocks is a matrix of shape (b, t, d), where b is the batch size, t is the sequence length, and d is the embedding dimension. These extracted features were then fed into a feedforward neural network, which outputs a vector out of dimension (b, c), where c is the number of classes. The predicted class label was determined by taking the *argmax* of the output vector.

In addition to BERT, we also explored the use of Gemma1.1 2B, a decoder-based architecture. This model was trained using an auto-aggressive approach with an attention mask on the output classification.

To compare with [1, 2, 3, 33, 5, 6], to utilize machine learning approaches, we used Term Frequency - Inverse Document Frequency (TF-IDF) as a feature extractor. TF-IDF is a statistical weighting technique that evaluates the importance of a term within a document by considering its frequency of occurrence, while also taking into account its rarity across the entire corpus. By balancing term frequency and inverse document frequency, TF-IDF provides a robust method for feature extraction and dimensionality reduction in text analysis. The extracted features from TF-IDF were inputted into Support Vector Machines (SVM) [21], logisitc regression [24], XGBoost [26], multi-layer perceptron (MLP), naive bayes multinomial [10], and random forest.

The dataset, comprising 13,389 samples, was divided into three subsets: 2,677 samples (20%) for testing, 1,071 samples (10%) for validation, and 9,640 samples (70%) for training. The training process for BERT and Gemma models was conducted on 2xT4 16GB GPUs. Specifically, BERT was trained for 7 epochs, while Gemma was trained for 1 epoch. In contrast, the classical methods (TF-IDF + <method>) were trained using the sci-kit learn library [13] on a Ryzen 7 12-core CPU. Notably, the training process for Gemma involved quantization-aware training with fp16 precision. Additionally, we employed LORA [14] to fit model weights on the GPU. The optimization process for both BERT and Gemma utilized AdamW [15] with a cosine learning rate scheduler. Furthermore, we implemented gradient

Method Name	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
TF-IDF + Random Forest [33]	78.5	75.2	82.1	78.6
TF-IDF + SVM [1, 3, 6]	79.2	76.5	83.5	80.0
TF-IDF + Logistic Regression [3]	79.8	77.1	84.2	80.6
TF-IDF + Naive Bayes Multinomial [1, 3, 6]	81.3	79.5	86.2	82.8
TF-IDF + MLP [2]	81.6	80.2	87.1	83.6
BiLSTM [4]	81.8	78.3	83.5	79.8
TF-IDF + XGB [6]	83.5	82.1	89.5	85.8
BERT	91.2	90.5	92.8	91.6
BERT - Top 3	96.1	95.8	97.3	96.5
BERT - Top 5	97.5	97.2	98.1	97.6
BERT - Top 10	98.5	98.3	99.1	98.7
Gemma1.1 2B	**92.0**	**91.5**	**93.2**	**92.3**

Table 1: Comparison of classification accuracy, precision, recall and f1 scores for different methods. The lower section presents our work.

check pointing [16] and gradient accumulation [17] to reduce GPU requirements, as the Adam optimizer maintains multiple copies of the weights.

5. Results & Discussion

We present the accuracy, precision, recall, and f1-score of each model on the test dataset, comprising 2,678 samples. For BERT, we report top-1, top-3, top-5, and top-10 accuracy, as some resumes can have multiple valid job titles. This is reflected in the top-x suggested titles according to a pre-chosen top-p probability.

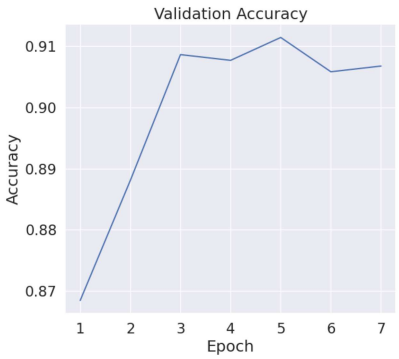
As shown in Table 1, attention-based models (Gemma, BERT) outperform their classical machine learning counterparts. The top-performing transformer-based model, Gemma1.1 2B, surpasses the top-performing classical model, TF-IDF + XGB, by 8.5%. This can be attributed to the ability of attention-based architectures to leverage the auto-aggressive nature of text, unlike TF-IDF, which relies solely on word frequency. Additionally, XGBoost performs well among classical machine learning methods, owing to its ensemble learning and regularization capabilities. Moreover, TF-IDF produces complex features that require a large space to represent, making MLP the best performer among classical methods, except for XGBoost.

Notably, all models were trained on only the first 300 words of the text. As illustrated in Figures 2c, 2d, nearly 70% of job titles are present in the first 300-500 words, and the probability of finding the job title in the text does not increase significantly beyond 500 words. This suggests that important information is typically found at the head of the resume.

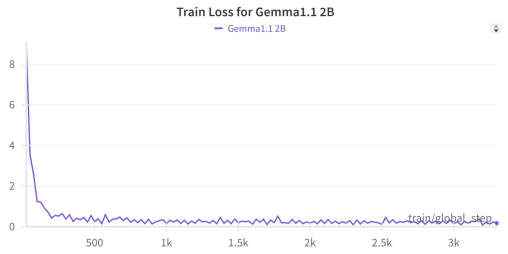
Figure 2b shows that Gemma training saturates quickly, around 250 steps out of 3900 steps (only 6%). This highlights the power of large language models in solving text classification tasks, particularly decoder-based architectures that exploit the auto-aggressive nature of input text. In contrast, Figure 2a reveals that, after the third epoch, the training loss decreases while the validation loss increases, indicating overfitting behavior in multi-epoch training for encoder-based architectures.

As evident from Table 1, increasing the number of output in BERT from top-1 to top-3 yields a 5.4% accuracy improvement, supporting our claim that some resumes can be appropriately labeled with multiple titles. This may be due to individuals holding multiple jobs or a single job being attributed to multiple titles. For example, the occupation of software engineer can be interchangeably referred to as frontend engineer, backend engineer, Python developer, full-stack engineer, and other related titles. Hence, considering the top-x samples would make sense.

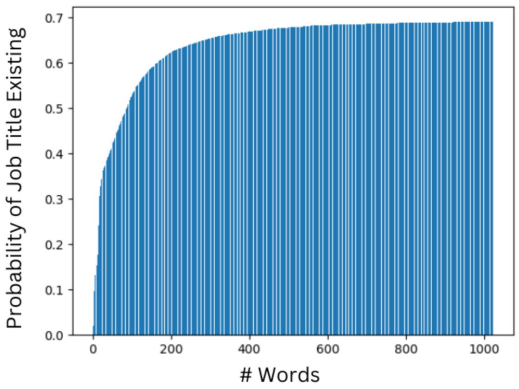
Finally, the high accuracy of our models can be attributed to the large size of our dataset, comprising 13,000 examples, and the diversity of job titles, with nearly twice as many titles as the largest existing dataset (Figure 2c). This ensures our models generalize well to unseen labels.



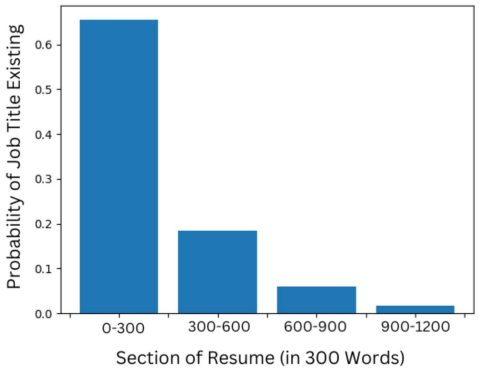
(a) BERT training loss and accuracy.



(b) Gemma training loss.



(c) Probability of title in text vs. the number of words in text.



(d) Probability of title in text vs. 300-word sections.

6. Conclusion

In this study, we addressed the significant challenges in resume classification within the online recruitment domain by leveraging a large-scale dataset and advanced Language Model Models (LLMs). Through meticulous dataset collection efforts, we assembled a comprehensive dataset comprising 13,389 resumes across 43 distinct categories. This extensive dataset enabled us to overcome limitations associated with small datasets and variability in resume formats, thereby enhancing the robustness and reliability of our classification models. Additionally, by employing state-of-the-art transformer models such as BERT and Gemma1.1 2B, we achieved remarkable improvements in classification accuracy, outperforming traditional machine learning approaches.

Our results demonstrate the efficacy of utilizing large datasets and advanced LLMs in improving the accuracy of resume classification tasks. The top-performing models achieved top-1 accuracy of 92% and a top-5 accuracy of 97.5%, underscoring the effectiveness of our approach in addressing the complexities of resume classification.

7. Future work

While our study represents a significant advancement in the field of resume classification, there are several avenues for future research and improvement. One promising direction is the collection of even larger datasets from a diverse range of sources, including social media platforms, job boards, and professional networking sites. Additionally, expanding the scope of labels and job titles covered in the dataset can further enhance the generalization capabilities of classification models. Furthermore, exploring innovative techniques for handling resume variability, such as adapting to different resume formats and incorporating multi-modal information, holds promise for improving classification accuracy and robustness. Overall, continued efforts in dataset collection, model development, and methodological innovation are essential for advancing the state-of-the-art in resume classification and enhancing the efficiency of online recruitment practices.

References

[1] Pal, R., Shaikh, S., Satpute, S., & Bhagwat, S. (2022). Resume classification using various machine learning algorithms. In ITM Web of Conferences (Vol. 44, p. 03011). EDP Sciences.

[2] Skondras, P., Zervas, P., & Tzimas, G. (2023). Generating Synthetic Resume Data with Large Language Models for Enhanced Job Description Classification. Future Internet, 15(11), 363.

[3] Ali, I., Mughal, N., Khand, Z. H., Ahmed, J., & Mujtaba, G. (2022). Resume classification system using natural language processing and machine learning techniques. Mehran University Research Journal Of Engineering & Technology, 41(1), 65-79.

[4] Jalili, A., Tabrizchi, H., Razmara, J., & Mosavi, A. (2024, January). BiLSTM for Resume Classification. In 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI) (pp. 000519-000524). IEEE.

[5] Nasser, S., Sreejith, C., & Irshad, M. (2018, July). Convolutional neural network with word embedding based approach for resume classification. In 2018 International Conference on Emerging Trends and Innovations In Engineering And Technological Research (ICETIETR) (pp. 1-6). IEEE.

[6] Ramraj, S., & Sivakumar, V. (2020, July). Real-Time Resume Classification System Using LinkedIn Profile Descriptions. In 2020 International Conference on Computational Intelligence for Smart Power System and Sustainable Energy (CISPSSE) (pp. 1-4). IEEE.

[7] DeepMind, G. T. G., Team, G., & DeepMind, G. (n.d.). Gemma: Open models based on Gemini Research and Technology. https://arxiv.org/html/2403.08295v1

[8] Kamath, U., Graham, K. L., & Emara, W. (2022). Bidirectional encoder representations from Transformers (Bert). Transformers for Machine Learning, 43–70. https://doi.org/10.1201/9781003170082-3.

[9] Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[10] Abbas, M., Memon, K.A., Jamali, A.A., Memon, S. and Ahmed, A., 2019. Multinomial Naive Bayes classification model for sentiment analysis. IJCSNS Int. J. Comput. Sci. Netw. Secur, 19(3), p.62.

[11] Steinwart, I. and Christmann, A., 2008. Support vector machines. Springer Science & Business Media.

[12] Loper, E. and Bird, S., 2002. Nltk: The natural language toolkit. arXiv preprint cs/0205028.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.

[14] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

[15] Loshchilov, I. and Hutter, F., 2017. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

[16] Chen, T., Xu, B., Zhang, C. and Guestrin, C., 2016. Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174.

[17] Lamy-Poirier, J., 2021. Layered gradient accumulation and modular pipeline parallelism: fast and efficient training of large language models. arXiv preprint arXiv:2106.02679.

[18] Lindner, J. (19 April 2024) Online recruitment industry statistics, GITNUX, Available at: https://gitnux.org/online-recruitment-industry

[19] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[20] Botev, A., De, S., Smith, S.L., Fernando, A., Muraru, G.C., Haroun, R., Berrada, L., Pascanu, R., Sessa, P.G., Dadashi, R. and Hussenot, L., 2024. RecurrentGemma: Moving Past Transformers for Efficient Open Language Models. arXiv preprint arXiv:2404.07839.

[21] Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B., 1998. Support vector machines. IEEE Intelligent Systems and their applications, 13(4), pp.18-28.

[22] Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

[23] Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[24] Weisberg, S., 2005. Applied linear regression (Vol. 528). John Wiley & Sons.

[25] Abubakar, H.D., Umar, M. and Bakale, M.A., 2022. Sentiment classification: Review of text vectorization methods: Bag of words, Tf-Idf, Word2vec and Doc2vec. SLU Journal of Science and Technology, 4(1 & 2), pp.27-33.

[26] Ramraj, S., Uzir, N., Sunil, R. and Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications, 9(40), pp.651-662.

[27] Flach, P.A. and Lachiche, N., 2004. Naive Bayesian classification of structured data. Machine learning, 57, pp.233-269.

[28] Paramita, D. (2020). Digitalization in talent acquisition: A case study of AI in recruitment.

[29] Javed, F., Luo, Q., McNair, M., Jacob, F., Zhao, M., & Kang, T. S. (2015, March). Carotene: A job title classification system for the online recruitment domain. In 2015 IEEE First International Conference on Big Data Computing Service and Applications (pp. 286-293). IEEE.

[30] Sachan, V. S., Katiyar, A., Somashekher, C., Chauhan, A. S., & Bhima, C. K. (2024). The Role Of Artificial Intelligence In HRM: Opportunities, Challenges, And Ethical Considerations. Educational Administration: Theory and Practice, 30(4), 7427-7435.

[31] Chen, D. (2022). Artificial Intelligence (AI) in Employee Selection: How Algorithm-Based Decision Aids Influence Recruiters' Decision-Making in Resume Screening (Doctoral dissertation, The University of Texas at Arlington).

[32] Dou, B., Zhu, Z., Merkurjev, E., Ke, L., Chen, L., Jiang, J., ... & Wei, G. W. (2023). Machine learning methods for small data challenges in molecular science. Chemical Reviews, 123(13), 8736-8780.

[33] Zaroor, A., Maree, M. and Sabha, M., 2017, November. JRC: a job post and resume classification system for online recruitment. In 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 780-787). IEEE.