RESEARCH ON APPLYING AI-POWERED NLP AND DATA VISUALIZATION TO SUPPORT CANDIDATE PROFILE PROCESSING

Do Phuc

University of Information Technology
Vietnam National University, Ho Chi Minh City

Abstract: Every year, reviewing academic titles is organized, and each candidate submits a detailed profile with numerous achievements in research and teaching. Handling this massive volume of profiles within a limited time frame is truly an intellectually and time-demanding task. In this report, we present the application of several AI-powered NLP advancements to support the profile review process. We automate information extraction from candidates' registration forms and utilize AI's natural language processing tools for data analysis. These include keyword extraction to identify research topics, author name recognition for foreign collaborators, article content summarization, and clustering articles by title to detect overlapping content. Additionally, we use visual representation tools to illustrate analysis results from multiple candidate profiles, highlighting each candidate's accomplishments. We hope that by developing this system, we can aid in reviewing candidate profiles and ease the challenges of profile evaluation.

Keywords: *candidate's profile, AI, keyword extraction, foreign author recognition, text summarization, article clustering for content comparison*

I Introduction

A candidate's profile typically includes two important reports: Form 1 - the registration form for review and recognition of eligibility for Professor or Associate Professor titles, which contains comprehensive information about the candidate, including personal information, lists of research projects, teaching materials, publications, and scientific reports. Form 3 is an overview report on the candidate's scientific research and teaching activities. Additionally, some councils may require a supplementary form with further details, including the author's name, paper title, and supporting evidence of the paper such as the journal, and impact factor…. Furthermore, reviewers may access detailed content, evidence, and full-text articles on the National Council's website.

The candidate's profile includes a large number of papers. Consequently, reading and understanding all the content in these profiles to detect duplication, uncover content themes, research directions, adherence to research topics, and identify international collaboration in paper article authorship is an arduous task that demands considerable time and intellectual effort.

With advancements in natural language processing (NLP) techniques powered by AI and large language models such as GPT 3.5 and 4.o, we now have many tools to support quick reading and comprehension of large text volumes in candidate profiles. In this paper, we extracted information from candidate profiles. We applied several AI-powered NLP techniques to perform tasks such as keyword extraction for quick understanding of document topics, extraction of foreign author names in publications to support quality assessment of international collaborations, text summarization to facilitate faster document reading, and clustering of texts to compare similarities and locate articles with overlapping contentRather than diving into the theoretical foundations of these methods, models, and algorithms, we focus on demonstrating the application of NLP technology powered by AI and visual representation to expedite profile profile processing. We have also applied this software to process real candidate profiles of 2024. This paper is organized as follows: 1) Introduction, 2) Key phrase Extraction, 3) Foreign Author Identification, 4) Text Summarization, 5) Clustering papers for similarity analysis 6) Using the Data Visualization techniques in the candidate's profiles 7) Conclusion and Future Directions.

II Key phrase Extraction

A key phrase (also known as a keyword phrase (Campos, 2018) (Florescu, 2017) consists of one or more words linked together and usually captures the main topic or content of a paragraph, document, or paper. Key phrases are often used in search, text analysis, and information extraction. They serve the following purposes

Summarizing Main Content: Key phrases highlight a paragraph's main points and core topics. Reviewers can quickly grasp an overview report or skim through a candidate's papers without reading the entire report or paper.

Enhancing Searchability: Key phrases improve search capabilities. Reviewers can find papers containing specific keyphrases. Papers sharing the same keyphrase can also be analyzed for similarity, helping to detect content overlap.

Aiding Text Analysis: Key phrases are used in text analysis algorithms to determine the frequency and appearance of important topics, thus supporting information extraction or deepening understanding of the text or paper.

Improving Connectivity: Key phrases can link different ideas or themes within a document, providing readers with a clearer picture of the overview report, paper, or scientific report.

Some AI-powered NLP models for extracting keywords from text include the following (Mehdi Allahyari, 2017):

2.1 Using TF-IDF (Term Frequency-Inverse Document Frequency): This method uses TF-IDF to identify words based on their frequency in the sentences of the text. This classic technique is simple to implement and can yield acceptable results.

2.2. Using Rake (Rapid Automatic Keyword Extraction): This method uses context patterns in the text to identify keywords. Rake does not require labeled data and analyzes word frequency and position to score phrases, then extracts phrases with high scores (Haque, 2018).

2.3. Deep Learning Models: Models based on LSTM/GRU sequences analyze semantic and contextual patterns in text. These models can be trained to identify key phrases based on labeled data. Transformer models (such as BERT and RoBERTa) are fine-tuned for key phrase extraction tasks. BERT (Jacob Devlin, 2019), for example, understands context from both directions (left and right) and can be used with specialized output layers to identify important words or phrases in the text.

We used AI-powered keyword extraction models to find key phrases in candidates’ papers. Below is the list of key phrases extracted from the candidate's paper titles.

Table 1. List of Articles Containing Key Phrases of a Candidate

#	Keyphrases
1	parallel proximal svm algorithm tailored [96]
2	local support vector regression [60][71]
3	local support vector machines [46][81]
4	support vector machines [46][49][53][65][72][81][92]
5	local svm algorithms [59]
6	training clustering models [113]
7	summarize vietnamese texts [113]
8	parallel learning algorithms [71]
9	large scale multi [69]
10	automatic learning algorithms [81]
11	classifying large datasets [59]
12	parallel learning [59][71]
13	large datasets [6][7][31][59][60][71]
14	class datasets [69][79]
15	ray images [97]
16	massive classification [16][49]

Based on the results in Table 1, we can see that the candidate used the key phrase “support vector machines” in numerous papers with paper codes [46][49][53][65][72][81][92]. These key phrases also allow us to understand the topics that the candidate has pursued throughout their work.

Table 2. Publication Timeline of Articles Containing the Key phrase "svm"

Publication years	The number of papers
2003	1 paper
2004	3 papers
2005	1 paper
2006	2 papers
2008	1 paper
2009	1 paper
2013	1 paper
2014	1 paper
2017	2 papers
2019	4 papers
2022	2 papers

In Table 2, we observe that the candidate published 1 paper in 2003, 3 papers in 2004, … and 2 in 2022. The period between the first and last publication of the key phrase “svm” is 19 years. This duration highlights the candidate's sustained commitment to the “svm” topic over an extended period.

We have also extracted key phrases from the candidate’s overview report. Table 3 provides the key phrases identified in the candidate’s overview report:

Table 3. List of some Key phrases in the Scientific Overview Report

In Vietnamese	In English
Nghiên cứu khoa học	1. Scientific research
Phân tích dữ liệu một chiều	2.Univariate data analysis
Phân tích dữ liệu nhiều chiều	3.Multivariate data analysis
An toàn người bệnh	4. Patient safety
Đại học Quốc gia Chonnam	5.Chonnam National University
Tiến sĩ và sau Tiến sĩ	6.Doctoral and post-doctoral
Độ chính xác thiết bị đo đường cầm tay	7. Accuracy of handheld glucose meters
Ảnh hưởng của Hematocrit (HCT)	8.Hematocrit (HCT) influence
Mô hình máy học	9.Machine learning model
Xử lý ảnh	10. Image processing
Outliers	11. Outliers
Cung dòng điện chuyển đổi	12.Transduced current curve
Biosensor	13. Biosensor
Mạng nơ-ron một lớp ẩn (SLFN)	14.Single hidden layer feedforward neural network (SLFN)

Based on the key phrases in Table 3, we can quickly understand the research directions of the candidate.

II Extraction of Foreign Authors Collaborating with the Candidate

In evaluating profiles for academic titles, it is essential to identify foreign authors collaborating with the candidate to assess the quality of collaboration. Several techniques are employed to distinguish foreign author names from Vietnamese author names, including:

3.1. Machine Learning Models Based on Linguistic Features:

Feature Analysis: This involves creating features based on name length, character frequency, word order (e.g., surname-first vs. first name-first), and distinctive characters (e.g., accents, circumflex, hyphens).

Classification Algorithms: Machine learning algorithms such as Random Forest, SVM, or Logistic Regression are used to train models to classify names as either Vietnamese or foreign based on these features.

3.2. Deep Learning Models with BiLSTM/CRF:

BiLSTM (Bidirectional Long Short-Term Memory):** A bidirectional RNN model that can analyze character sequences from both directions, effectively recognizing names.

CRF (Conditional Random Field):** Combined with BiLSTM to recognize name sequences in context, increasing labeling accuracy.

3.3. Transformer-Based Model (BERT):**

Pre-trained Language Models: Models such as BERT (Jacob Devlin, 2019) can be fine-tuned to recognize named entities (NER – Named Entity Recognition). By using a labeled dataset, BERT can be trained to differentiate Vietnamese author names from foreign ones.

Fine-tuning: Tools like Hugging Face Transformers are used to fine-tune the BERT model on labeled data with person names marked as either Vietnamese or foreign.

In this study, we used the BERT NER model (Jacob Devlin, 2019) (Truong H. V. Phan, 2021) to identify foreign author names. This model utilizes BERT with an added Softmax layer for classification. The training dataset includes a dictionary of common Vietnamese names.

The program automatically identified foreign author names in a candidate’s profile, as shown below. Table 4 displays the list of foreign authors collaborating with candidates.

Table 4. List of Foreign Authors Collaborating with the Candidate

#	The foreign authors’ names
1	Heiko A. Schmidt
2	Arndt Von Haeseler
3	Andrés Varón
4	Ward C Wheeler
5	Michael Tillich
6	Katrin Schulerowitz
7	Uwe G Maier
8	Christian Schmitz Linneweber
9	Ward C. Wheeler
10	Bart Hazes
11	Tze Min Teo
12	Tomáš Flouri
13	Alex
14	Ros Stamatakis
15	Hanon Mcshea
16	Joanna Masel
17	Jennifer Eleanor James
18	Robert Lanfear

Table 4 shows the list of foreign authors collaborating with candidate. We then applied additional techniques to assess the reputation of these foreign authors.

IV Text Summarization

Text summarization (Mehdi Allahyari, 2017) (Mahak Gambhir, 2022) is a crucial technique in natural language processing (NLP). The goal is to create a concise text that retains the main points of the original document. There are two main approaches: extracting key sentences from the text (extractive summarization) and generating a summary in new words (abstractive summarization). Below are some common techniques in NLP for text summarization:

4.1. Extractive Summarization

This approach selects the most important sentences or paragraphs from the original document to form a summary. Common methods include:

- TF-IDF: Identifies sentences with the most important words based on their TF-IDF scores.

-PageRank, TextRank: Appling algorithms similar to PageRank to evaluate sentence importance based on the connections between sentences.

- LSTM-based Ranking: Using LSTM to predict and rank significant sentences.

4.2. Abstractive Summarization:

Generating a new summary with original sentences that still capture the key points from the original text, similar to how a human would summarize. Common models include:

- Seq2Seq Models: Using an encoder-decoder model with LSTM or GRU layers to transform the original text into a summary.

- Transformer-based Models (BART, T5): Transformer-based models like BART (Bidirectional and Auto-Regressive Transformers) or T5 (Text-to-Text Transfer Transformer) are trained to produce accurate and natural summaries.

- GPT: Using versions of GPT (Generative Pre-trained Transformer) to automatically create shorter texts that maintain semantic similarity with the original document.

4.3. Machine Learning and Deep Learning-Based Summarization: Some popular models include:

- Supervised Machine Learning Models: Using classification models to predict which sentences should be included in the summary, based on features such as sentence length, position, word frequency, and more.

- Fine-tuning Pre-trained Models: Fine-tunes models like BERT and BART on labeled data to generate customized summaries.

- LSTM or GRU Encoders: Using LSTM or GRU as an encoder to read and process the original text, then decodes it to produce a concise summary.

We used ChatGPT 4.o to generate a summary of one of the candidate’s papers. The prompt to ChatGPT was: “Generate the summary of the following paper.” The result was as follows:

The paper uses General Mapping Convergence Framework," presents an innovative approach to Single Class Classification (SCC), which involves distinguishing one class of data (positive) from other classes when negative data is scarce or difficult to obtain. This is common in situations where only positive and unlabeled data are available, making traditional classification methods ineffective due to class imbalance. The authors propose a method that integrates Extreme Learning Machine (ELM)—a rapid learning algorithm—with a mapping convergence algorithm inspired by Support Vector Machine (SVM) frameworks. This combined approach seeks to enhance classification performance while maintaining efficient learning speed and operational efficiency. The paper discusses how conventional SCC techniques, such as probability density estimation and boundary-defining methods, face challenges in high-dimensional data environments

We used this summary to support a quick reading of the candidate’s paper and to proceed with further analysis.

V Clustering papers for similarity analysis

Clustering is a technique used to group documents based on the similarity of their content. Since only the titles of the articles are available in the candidate's profile, we used the BERT model (Omar Galal, 2022) to create feature vectors for these titles. Then, clustering models such as k-means were used to group the feature vectors for these titles. The clustering results for a candidate's article titles are as follows.

Table 5. clustering results for a candidate XXXX's article titles

Paper IDs of Cluster 2:
[1][6][7][8][9][10][11][12][17][19][22][26][27][29][30][31][34][37][38][39][41][42][44][47][48][52][53][56][58][64]
Paper IDs of Cluster 0
[2][3][4][5][13][15][16][18][23][24][32][35][43][45][46][49][51][60][63]
Paper IDs of Cluster 1
[14][20][21][25][28][33][36][40][50][54][55][57][59][61][62]

Based on the results above, we identified groups of articles within a topic (cluster) for the candidate, enabling us to examine potential duplicates and compare the content of two articles

(Tham Vo Thi Hong, 2019). We selected the parameter k as the number of research directions of the candidate

Similarity Comparison Based on Abstract Keyphrases

We extracted key phrases from each article's abstract, then calculated the number of common key phrases between pairs of abstracts. If two abstracts share more than 15 keyphrases, we investigate the similarity between the two corresponding articles. Below is a comparison result for the key phrases of articles 15 and 16 for a candidate.

Common Key phrases of the Abstracts of paper 15 and paper 16:

Common key phrases: ['transduced current curve', 'type electrochemical biosensors', 'important factor', 'oxidase reaction', 'handheld devices', 'nonlinear methods', 'glucose measurements', 'hematocrit estimation', 'biosensors', 'glucose', 'hematocrit', 'hct', 'words', 'strip', 'produced', 'paper', 'key', 'abstract']

Number of common key phrases: 18

With a similarity threshold of over 15 key phrases, the program indicates the similarity between the abstract of Paper 15 and Paper 16. This is the first step of the overlap analysis in two papers.

VI Using The Data Visualization Techniques In The Candidate's Profiles

We also used data visualization techniques to represent certain information in the candidate’s profile visually.

Visualization of Publication Progress

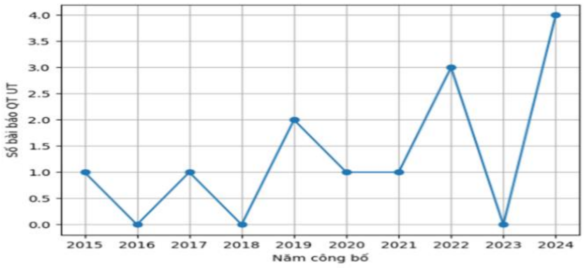

Figure 1. Publication Progress of Scientific Papers by the Candidate after Attaining Associate Professor title

Figure 1 shows a chart with the years of publication on the x-axis and the number of papers published each year on the y-axis. From Figure 1, we can observe the Candidate’s significant effort in the final year (2024), with a total of 4 papers published.

Visualization of Paper Citations

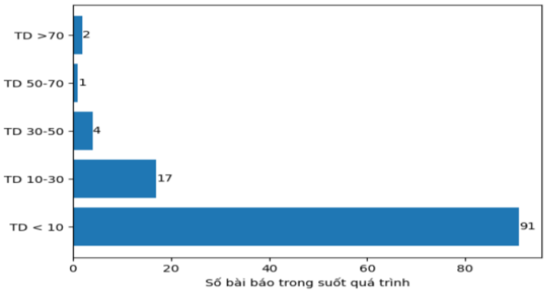

Figure 2. Chart of the Candidate’s Papers and Citation Counts

The chart in Figure 2 displays the papers along the x-axis and the number of citations per paper on the y-axis. Through this chart, we can quickly assess the impact of the candidate's publications. It shows that only two of the candidate's papers have more than 70 citations.

Visualization of Authors per Paper

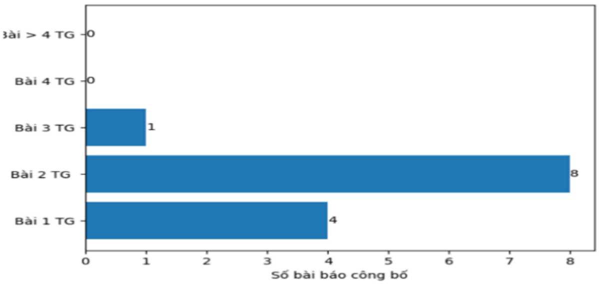

Figure 3. Chart of Papers and Number of Authors

The chart in Figure 3 shows the number of papers on the x-axis and the number of authors per paper on the y-axis. From the chart, we can observe that this candidate has four papers authored solely by themselves, indicating a strong ability to write papers independently.

Visualization of Journals and Number of Candidates Publishing in Each Journal

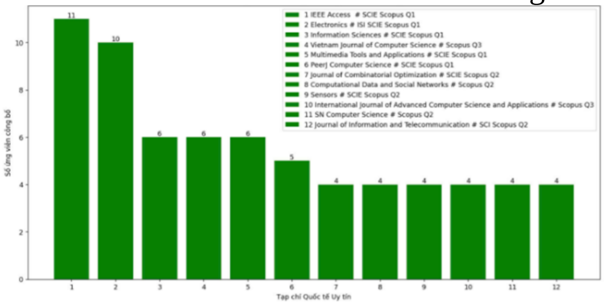

Figure 4. Chart of Journals and Number of Candidates Publishing in Each Journal

The chart in Figure 4 displays the prestigious international journals on the x-axis and the number of Candidates who have published in each journal on the y-axis. From the chart, we can see that the journal named IEEE Access has 15 candidates who have published articles. This indicates that it is a popular choice among many candidates.

Chart Surveying Keyphrases and the Number of Candidates Using Keyphrases

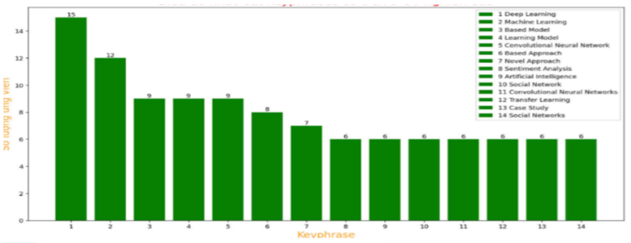

Figure 5. Chart Surveying Key phrases and the Number of Candidates Using Key phrases

The chart in Figure 5 shows the key phrases on the x-axis and the number of candidates using those keyphrases in their papers on the y-axis. From the chart in Figure 5, we observe that the keyphrase “deep learning” has been used by 15 candidates. Thus, the majority of candidates are focusing their research on deep learning models. This is a current trend in information technology.

These visualization tools help evaluators quickly grasp the accomplishments of the candidates.

VII Conclusion and Future Directions

The candidate's profile contains a vast amount of information regarding their scientific research outcomes. These profiles often have numerous papers that need to be read and processed within a s short time. Reading a large volume of papers in a short time is indeed a challenge. In this paper, we present the application of several AI-powered NLP such as keyword extraction, extraction of foreign authors' names, content summarization, text clustering, and visualization to assist evaluators in quickly reading and understanding the papers and scientific reports within the candidates' profiles. We also present results based on real profiles. We are continuing to refine and develop the software with the hope of creating a tool that aids in reading and comprehending candidate profiles.

Acknowledgment.

The author sincerely thanks Prof. DrSc. Hoàng Văn Kiếm for his assistance in completing this paper.

REFERENCES

1. Ashish Vaswani et al. "Attention is All You Need," Neural Information Processing, June 2017.
2. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," https://doi.org/10.48550/arXiv.1810.04805, 2019.
3. Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., Jatowt, A. "YAKE! Collection-independent Automatic Keyword Extractor," Proceedings of the 41st European Conference on Information Retrieval (ECIR), 2018.
4. Florescu, C., Caragea, C. "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents," Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2017.
5. Mahak Gambhir, Vishal Gupta. "Deep Learning-based Extractive Text Summarization with Word-level Attention Mechanism," Multimedia Tools and Applications, Volume 81, Issue 15, 2022.
6. Mehdi Allahyari et al. "Text Summarization Techniques: A Brief Survey," https://doi.org/10.48550/arXiv.1707.02268, 2017.
7. Mozammel Haque. "Automatic Keyword Extraction from Bengali Text Using Improved RAKE Approach," 21st International Conference of Computer and Information Technology (ICCIT), 2018.
8. Omar Galal, Ahmed H. Abdel-Gawad, Mona Farouk. "Rethinking of BERT Sentence Embedding for Text Classification," Neural Computing and Applications, Volume 36, pages 20245–20258, 2024.
9. Tham Vo Thi Hong, Phuc Do. "Comparing Two Models of Document Similarity Search over a Text Stream of Articles from Online News Sites," Intelligent Computing and Optimization: Proceedings of the 2nd International Conference on Intelligent Computing and Optimization 2019 (ICO 2019).
10. [10] Truong H. V. Phan, Phuc Do. "NER2QUES: Combining Named Entity Recognition and Sequence to Sequence to Automatically Generate Vietnamese Questions," Neural Computing and Applications, Volume 34, pages 1593–1612, 2022.

NGHIÊN CỨU ỨNG DỤNG AI NLP VÀ BIỂU DIỄN TRỰC QUAN ĐỂ HỖ TRỢ XỬ LÝ HỒ SƠ ỨNG VIÊN

Tóm tắt: Hàng năm công tác xét duyệt chức danh khoa học đều được tổ chức. Các ứng viên đều có hồ sơ rất chi tiết với nhiều thành tích trong nghiên cứu, giảng dạy. Việc xử lý khối lượng đồ sộ hồ sơ này trong khoảng thời gian ngắn thực sự là một công việc đòi hỏi thách thức về trí tuệ và thời gian. Trong báo cáo này, chúng tôi trình bày việc áp dụng một số thành tựu của NLP AI để hỗ trợ công tác xét duyệt hồ sơ. Chúng tôi tự động rút trích thông tin trong bản đăng ký của ứng viên và ứng dụng một số công cụ xử lý xử lý ngôn ngữ tự nhiên của AI vào công việc phân tích dữ liệu như: rút trích từ khóa thể hiện chủ đề nghiên cứu, nhận dạng tên tác giả nước ngoài mà ứng viên có cộng tác, tóm tắt nội dung các bài báo, gom cụm bài báo theo tên bài báo để phát hiện các bài báo có nội dung trùng lặp…..Bên cạnh đó, chúng tôi sử dụng các công cụ biểu diễn trực quan các kết quả phân tích nhiều hồ sơ ứng viên để thể hiện các kết quả của ứng viên. Chúng tôi hy vọng với việc xây dựng hệ thống này có thể hỗ trợ công tác đọc hồ sơ ứng viên và giảm nhẹ khó khăn trong công tác thẩm định hồ sơ ứng viên.

Từ khóa: hồ sơ ứng viên, AI, rút trích từ khóa, rút trích tên tác giả nước ngoài, tạo tóm tắt văn bản, gom cụm bài báo so sánh nội dung bài báo

Professor Dr. Do Phuc is currently working at the University of Information Technology, VNU-HCM. His research areas include bioinformatics, text processing, artificial intelligence, machine learning, deep learning, natural language processing, knowledge graphs, question answering, verification, and big data processing. Professor Do Phuc has participated in numerous research projects across various fields, including bioinformatics, text processing, natural language processing, text understanding systems, and reasoning over knowledge graphs. He has published more than 60 scientific papers in prestigious international and national journals. He is the author of 8 reference books and textbooks in the fields of bioinformatics, computer science, and applications.

You can contact him via email: phucdo@uit.edu.vn.

Corresponding author: Prof. Dr. Phuc Do
Email: phucdo@uit.edu.vn