

SkillRec: A Data-Driven Approach to Job Skill Recommendation for Career Insights

Xiang Qian Ong

Information Systems Technology and Design Pillar
Singapore University of Technology and Design
Singapore
xiangqian_ong@alumni.sutd.edu.sg

Kwan Hui Lim

Information Systems Technology and Design Pillar
Singapore University of Technology and Design
Singapore
kwanhui_lim@sutd.edu.sg

Abstract—Understanding the skill sets and knowledge required for any career is of utmost importance, but it is increasingly challenging in today’s dynamic world with rapid changes in terms of the tools and techniques used. Thus, it is especially important to be able to accurately identify the required skill sets for any job for better career insights and development. In this paper, we propose and develop the Skill Recommendation (SkillRec) system for recommending the relevant job skills required for a given job based on the job title. SkillRec collects and identify the skill set required for a job based on the job descriptions published by companies hiring for these roles. In addition to the data collection and pre-processing capabilities, SkillRec also utilises word/sentence embedding techniques for job title representation, alongside a feed-forward neural network for job skill recommendation based on the job title representation. Based on our preliminary experiments on a dataset of 6,000 job titles and descriptions, SkillRec shows a promising performance in terms of accuracy and F1-score.

Index Terms—Recommendation Systems, Skill Recommendation, Occupational Analysis, Natural Language Processing, Neural Networks

I. INTRODUCTION

Technological advancements and industrial changes are constantly evolving at a rapid pace [1]. With this rapid change, new job scopes are constantly being generated at companies along with a dynamic list of skill sets to meet these ever-growing changes. As such, individuals that are looking to change their career or seeking a new job often face the problem of skill mismatch due to those constant changes. This problem is further highlighted by the recent COVID-19 pandemic which has caused drastic changes in many industries, such as aviation and tourism. During this period, there was a spike in the number of retrenched workers seeking to find employment in other fields [2] and potentially facing skill mismatch.

On top of that, jobs are tightly coupled with skill sets, which are essential for fulfilling the job requirements. Individuals who are transiting into a new career and industry would need to conduct extensive research or seek advice on the latest job requirements and the relevant skills. This process is commonly performed to improve the chances of applicants securing a new job, especially in a new industry. This is necessary due to the huge barrier of entry for individuals transiting over to new jobs or new industries, given their existing skill sets. Different industry fields and even roles within the same industry require

different sets of skills to carry out tasks associated with the job. Likewise, hiring managers also look out for those skill sets when assessing suitable candidates for an open role. As such, identifying the skill sets required for a given job is especially important for potential job seekers, even for those staying in the same industry.

A. Related Work

In terms of academic research, this problem and research area have also garnered keen interest in recent years due to these various developments [1], [3]–[7]. For example, there are various streams of research that investigate diverse but related topics such as: (i) recommending job candidates to companies based on their suitability [8]–[10]; (ii) recommending jobs to candidates based on the candidates skill sets and career history [11], [12]; (iii) modelling the career trajectories of individuals [13], [14]; (iv) predicting the future career transitions and movements of workers [15], [16].

Among these works, the most closely related works to our research are those on job skill recommendations [17]–[19] and those using various embedding techniques [4], [20]–[22] for occupational analysis and mining. For example, Dave et al. proposed a representation learning model using job transitions, job-skills and skill co-occurrence networks for recommending both jobs and skills [17]. Others have used reinforcement learning in the form of a Deep-Q Network for skill recommendation and better interpretation [19]. Liu et al. developed a bi-directional language model for contextualised job tile embeddings that was demonstrated on a downstream occupational named entity recognition task [4]. Skills2Graph [21] used FastText embeddings to better maintain semantic similarity between taxonomic components of job postings. Also closely related to our research are various works that have developed systems and implemented approaches for related occupational analysis and mining tasks [20], [23], [24].

B. Motivation and Problem Statement

The fundamental problem is that identifying the relevant skill sets required for a particular job title is both time-consuming and a complex process. However, this process is essential to job seekers looking to enter new jobs and yet very challenging as they are new to both the field and industry.

The problem can be broken down into two sub-problems: the lack of knowledge to learn the relevant skills and the lack of proficiency in the relevant skills, which will be elaborated next.

1) *Lack of knowledge to learn the relevant skills:* To learn the relevant skills required for the particular job, there is a need to understand the current job workflow, the tools that are used on the job and knowledge in that field of interest. The root of the problem is the lack of domain expertise in the new industry to truly understand the skill sets required for the job [25].

There are various different ways in which the problem can be mitigated. The first is to seek advice from experts who have worked in the field for a long time. They understand the job requirements and the necessary skill sets needed, from their extensive years of experience. Alternatively, governments' skills recommendations serve as a good indicator of possible general relevant skill sets that are required in the industry for the near future [26], [27]. Lastly, individuals could perform online research on the industry, given that there is an abundance of articles and forums that share insights and recommendations.

This step is crucial as it serves as the basis for the next step. Only after knowing the skills that need to be learned, then the individual can embark on the next step.

2) *Lack of proficiency in the relevant skills:* After identifying the required skill sets to transit to a new industry or role, it is crucial for individuals to have access to resources to learn those skills. Proficiency in these relevant skills is an essential requirement that leads to successful employment and subsequent progression. The resources could come in many different forms, both physical and digital, which we discuss next.

A classroom setting is a common structure whereby individuals will attend classes and undergo a structured curriculum to gain knowledge and experience from the instructors. At the end of such courses, there are formal certificates that act as proof of completion of the course. Classroom formats could also be virtual, ranging from online courses such as Udemy and Coursera. Similarly, these classes provide both the content and practical lessons, with the added bonus of online certificates upon completion. Openly available resources are unstructured curricula that could comprise different resources from both online and offline. This gives the individual the flexibility of time and learning opportunity but typically do not result in a formal certification upon completion.

C. Problem Definition

In this paper, our main focus is on the first sub-problem, which is the lack of knowledge to learn the relevant skills required for a job role. This problem is significantly more crucial than the second sub-problem, as the former's outcome will determine the skill sets that an individual is required to learn. Failing at this step will render useless any future efforts at both a industry transition or career progression. In addition, there are multiple open-source and affordable resources online to address the problem of a lack of proficiency, thus this second

sub-problem is thus not as crucial as compared to the first sub-problem that we focus on. More formally, we investigate the problem of identifying and recommending the required job skill sets required for a role given the job title.

D. Main Contributions

The current ways to address the problems are mainly human-driven insights, such as seeking the advice of current practitioners in the new industry or through self-reading of openly available resources. However, we could tap into data-driven insights that perform the analysis and recommend the right skill sets to pick up for the industry or role we are interested in. This could reduce the time taken or the cost incurred in the process.

In this paper, we make the following contributions:

- 1) We investigate the problem of job skill recommendation, which is to recommend a list of required skill sets given role in the form of a job title.
- 2) To address the above problem, we propose and develop the Skill Recommendation (SkillRec) system. SkillRec comprises the capability to collect openly available job listing information, process these data to extract job title and required skill sets, and subsequently recommend skill sets given a job title.
- 3) As part of SkillRec, we also develop two approaches to skill recommendation, utilising word/sentence embedding techniques (using BERT and FastText) for job title representation, before inputting this job title representation to a simple but effective feed-forward neural network for job skill recommendation.
- 4) Using a dataset of 6,000 job listings, we perform preliminary experiments to evaluate the performance of our proposed SkillRec system. Our preliminary results show promising performance of SkillRec in terms of accuracy and F1-score.

II. PROPOSED SKILL RECOMMENDATION SYSTEM

Our proposed Skill Recommendation (SkillRec) system comprises four main components, including a web collection component, data pre-processing component, job title representation component and job skill recommendation component. Figure 1 shows an overview of our proposed system. Next, we provide a brief summary of each component, followed by a more detail description in the later sections.

- 1) **Web Collection Component.** This component retrieves job listings from various web sources, such as online job posting platforms.
- 2) **Data Pre-processing Component.** This component performs the necessary data pre-processing and cleaning to extract relevant information from the job descriptions, such as the job title, job description, required skills, company name, etc.
- 3) **Job Title Representation Component.** This component utilises word/sentence embedding techniques, such as BERT and FastText, to model job titles in term of its vector representations.

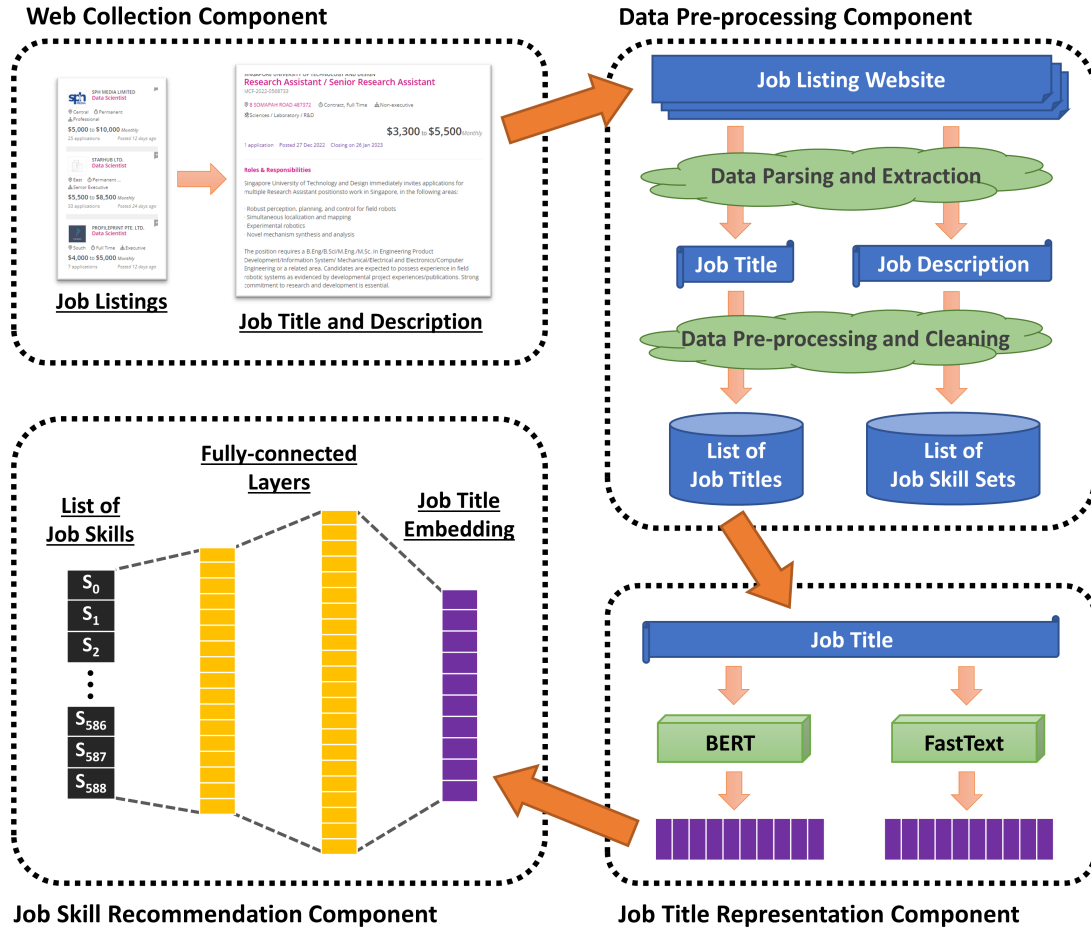


Fig. 1. System Architecture of Our Proposed Skill Recommendation (SkillRec) System

4) Job Skill Recommendation Component. The job title representation from the previous step is then fed into this component, which comprises a feed-forward neural network that outputs a set of recommended skills given this job title.

A. Web Collection Component

The web collection component retrieves various types of web-based information such as job descriptions from popular online job posting platforms, as well as skill sets from massive open online courses platforms.

The job description data was retrieved over a period of six months at a weekly interval to ensure that the dataset is kept up-to-date. A total of 6,000 job descriptions were retrieved from three popular online job posting platform.

Massive open online courses platforms provide a good source of information on popular educational courses and the skill sets they aim to impart. Using two popular massive open online courses platforms, we collected and identified 589 unique skill sets that are relevant for this work.

B. Data Pre-processing Component

The data pre-processing component works on the job title and job description data collected by the web collection component. This component first parses the collected webpage to extract the relevant information, i.e., the job title and job description. Thereafter, the standard text pre-processing steps, such as conversion to lowercase, removal of non-textual symbols, tokenisation, are performed on these two fields. The skill sets for a job listing is derived from its corresponding job description based on an exact match of a mentioned skill with our earlier list of 589 unique skill sets.

C. Job Title Representation Component

As our main task is to determine the required skill sets for a given job title, we proceed to model the job titles using various embedding techniques. In this study, we experiment with two popular word/sentence representation techniques:

- BERT Embedding [28]. The Bidirectional Encoder Representation of Transformer (BERT) is a transformer-based language model comprising multiple layers of bidirectional Transformer encoders, along with self-attention

mechanism. BERT is then pre-trained with the Masked Language Model task of predicting hidden tokens using other non-hidden tokens, and/or Next Sentence Prediction task of predicting the next sentence given an input sentence.

- FastText Embedding [29], [30]. FastText builds upon the skip-gram approach that is popularly used by the Word2Vec model [31], [32]. Using FastText, each word is represented by a bag-of-character n-grams, and in turn each character n-grams is represented by a vector. The vector representation of each word is then derived from the sum of these character n-grams vector representations.

BERT was chosen due to its strong performance in various language modelling tasks, such as text classification, question answering, among others. In addition, BERT has been frequently used in a variety of other recommendation and prediction tasks, such as location prediction [33], [34] and tourism recommendation [35], [36].

After extracting the job title in the previous step, either BERT or FastText is then used to obtain the vector representation of the job title. This job title representation is subsequently used as input to our neural network for job skill recommendation, which we describe next.

D. Job Skill Recommendation Component

Using the vector representation of the job title from the previous step, this job title representation is then fed into a feed-forward neural network for job skill recommendation based on the job title. We utilise a feed-forward neural network with two fully-connected hidden layer, with the first hidden layer comprising 1280 nodes, the second hidden layer comprising 640 nodes and a final output layer of 589 nodes corresponding to the 589 unique skill sets. We employ Sigmoid as our activation function, AdamW as our optimizer and a dropout of 0.5 to reduce overfitting.

III. EXPERIMENTS AND RESULTS

Using our collected dataset, we use 80% for training our model as described in Section II and use the remaining 20% for evaluation. For evaluation purposes, we use the standard metrics of accuracy and F1-score, based on the recommended job skill against the ground truth job skill in the original job description.

We experimented with the following algorithms and baselines:

- Bag-of-Words. A classical bag-of-words model that is traditionally used for classification tasks involving text-based data. In our application, we utilise a frequency-based approach to model the frequency of which a particular skill is mentioned in relation to a specific job. This model is used as a simple baseline to compare our two proposed approaches.
- FastText+NN. As described earlier in our SkillRec system, this model uses the FastText embedding for job title representation (Section II-C), followed by a neural network (NN) for job skill recommendation (Section II-D).

- BERT+NN. Similar to the earlier FastText+NN model, this model uses the BERT embedding for job title representation (Section II-C), followed by the same neural network (NN) for job skill recommendation (Section II-D).

TABLE I
EXPERIMENTAL RESULTS

Algorithm	Accuracy	F1-score
Bag-of-Words	0.4768	0.2887
FastText+NN	0.9728	0.4931
BERT+NN	0.9870	0.4973

Table I shows the average results of our experiments for the three models in terms of accuracy and F1-score. The results show that our proposed SkillRec system using BERT+NN is the best performing model, with FastText+NN being a extremely close second. BERT has been shown to perform well for a range of natural language processing tasks, from text classification to question answering. Similarly for this task of occupational skill recommendation, BERT as well as FastText perform well with the add-on of a straightforward neural network. Unsurprisingly, Bag-of-Words performs the worst by a large margin due to the simple job title representation approach that is unable to capture similar job titles obtained.

TABLE II
EXAMPLE JOB TITLE AND RELEVANT SKILLS

Job Title	Related Jobs and Skills
developer	java, python, fullstack, android, software
software	stack, java, python, fullstack, android
chef	pastry, kitchen, hair, chemist, accountant
senior director	director, partner, pacific, planner, strategy
software engineer	software, engineer, python, java, embedded

In addition to the experiments measuring accuracy and F1-score of the job skill recommendations, we also performed a qualitative check on the job title representations learnt and the similar job titles and skill sets based on these representations. Table II shows a sample of some job titles and the related job titles and skill sets. This simple analysis shows the effectiveness of our approach as occupations such as developer and software engineer results in relevant skills like Java, Python, fullstack, embedded, etc. Similarly, non-technical roles such as senior director also resulted in similar roles like partner and planner and relevant skills like strategy. For certain roles like chef, the results are more mixed with relevant terms like pastry and kitchen being mentioned but also non-relevant ones like chemist and accountant. This issue highlights a limitation due to our preliminary experiments and limited dataset, which contains more technical roles than non-technical ones.

IV. CONCLUSION AND FUTURE WORK

In this paper, we proposed the SkillRec system to address the problem of job skill recommendation given a job title. SkillRec comprises four main components for web collection, data pre-processing, job title representation and job skill

recommendation. In particular, our preliminary experiments show that an approach that combines job title representation using BERT, along with job skill recommendation using a straightforward feed-forward neural network works effectively in terms of accuracy and F1-score.

Future enhancements to SkillRec can involve a more dynamic recommendation of job skills to account for new skill sets that may emerge in future or existing skill sets that evolve over time. Similarly, we can also include more context in addition to the job titles, such as the industry, country, job applicant details, etc, for making more contextualised and personalised job skill recommendations.

V. ACKNOWLEDGMENT

This research is funded in part by the Singapore University of Technology and Design under grant RS-MEFAI-00005-R0201.

REFERENCES

- [1] I. Khaouja, I. Kassou, and M. Ghogho, "A survey on skill identification from online job ads," *IEEE Access*, vol. 9, pp. 118 134–118 153, 2021.
- [2] W. E. Forum, "The future of jobs report 2020," *World Economic Forum*, 2020.
- [3] A. Guhnani and H. Misra, "Implicit skills extraction using document embedding and its use in job recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13 286–13 293.
- [4] J. Liu, Y. C. Ng, Z. Gui, T. Singhal, L. Blessing, K. L. Wood, and K. H. Lim, "Title2vec: A contextual job title embedding for occupational named entity recognition and other applications," *Journal of Big Data*, vol. 9, no. 1, pp. 1–16, 2022.
- [5] Z. Ao, G. Horváth, C. Sheng, Y. Song, and Y. Sun, "Skill requirements in job advertisements: A comparison of skill-categorization methods based on wage regressions," *Information Processing & Management*, vol. 60, no. 2, p. 103185, 2023.
- [6] A. Barducci, S. Iannaccone, V. La Gatta, V. Moscato, G. Sperli, and S. Zavota, "An end-to-end framework for information extraction from italian resumes," *Expert Systems with Applications*, vol. 210, p. 118487, 2022.
- [7] J. Liu, Y. C. Ng, K. L. Wood, and K. H. Lim, "Ipod: a large-scale industrial and professional occupation dataset," in *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 2020, pp. 323–328.
- [8] C. Zhu, H. Zhu, H. Xiong, C. Ma, F. Xie, P. Ding, and P. Li, "Person-job fit: Adapting the right talent for the right job with joint representation learning," *ACM Transactions on Management Information Systems (TMIS)*, vol. 9, no. 3, p. 12, 2018.
- [9] D. Shen, H. Zhu, C. Zhu, T. Xu, C. Ma, and H. Xiong, "A joint learning approach to intelligent job interview assessment," in *IJCAI*, 2018, pp. 3542–3548.
- [10] C. Qin, H. Zhu, T. Xu, C. Zhu, L. Jiang, E. Chen, and H. Xiong, "Enhancing person-job fit for talent recruitment: An ability-aware neural network approach," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 25–34.
- [11] J. Malinowski, T. Keim, O. Wendt, and T. Weitzel, "Matching people and jobs: A bilateral recommendation approach," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, vol. 6, 2006, pp. 137c–137c.
- [12] Y. Zhang, C. Yang, and Z. Niu, "A research of job recommendation system based on collaborative filtering," in *2014 Seventh International Symposium on Computational Intelligence and Design*, vol. 1, 2014, pp. 533–538.
- [13] Y. Liu, L. Zhang, L. Nie, Y. Yan, and D. S. Rosenblum, "Fortune teller: predicting your career path," in *Thirtieth AAAI conference on artificial intelligence*, 2016.
- [14] D. Mimno and A. McCallum, "Modeling career path trajectories," 2008.
- [15] C. James, L. Pappalardo, A. Sirbu, and F. Simini, "Prediction of next career moves from scientific profiles," *arXiv preprint arXiv:1802.04830*, 2018.
- [16] Y. Yang, D.-C. Zhan, and Y. Jiang, "Which one will be next? an analysis of talent demission," *Proceedings of the 1st International Workshop on Organizational Behavior and Talent Analytics*, 2018.
- [17] V. S. Dave, B. Zhang, M. Al Hasan, K. AlJadda, and M. Korayem, "A combined representation learning approach for better job and skill recommendation," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018, pp. 1997–2005.
- [18] A. Maurya and R. Telang, "Bayesian multi-view models for member-job matching and personalized skill recommendations," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, 2017, pp. 1193–1202.
- [19] Y. Sun, F. Zhuang, H. Zhu, Q. He, and H. Xiong, "Cost-effective and interpretable job skill recommendation with deep reinforcement learning," in *Proceedings of the Web Conference 2021*, 2021, pp. 3827–3838.
- [20] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzananza, and A. Seveso, "Skills2graph: Processing million job ads to face the job skill mismatch problem," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.
- [21] —, "Skills2job: A recommender system that encodes job offer embeddings on graph databases," *Applied Soft Computing*, vol. 101, p. 107049, 2021.
- [22] C. M. Jaramillo, P. Squires, H. G. Kaufman, A. M. Da Silva, and J. Togelius, "Word embedding for job market spatial representation: tracking changes and predicting skills demand," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 5713–5715.
- [23] X. J. S. Ashok, E.-P. Lim, and P. K. Prasetyo, "Jobsense: A data-driven career knowledge exploration framework and system," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2018, pp. 1411–1416.
- [24] C. Qin, H. Zhu, C. Zhu, T. Xu, F. Zhuang, C. Ma, J. Zhang, and H. Xiong, "Duerquiz: A personalized question recommender system for intelligent job interview," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2165–2173.
- [25] P. Restrepo, "Skill mismatch and structural unemployment," *MIT Working Paper*, 2015.
- [26] A. Felstead, D. Gallie, F. Green, and Y. Zhou, *Skills at Work in Britain, 1986 to 2006*. ESRC Centre on Skills, Knowledge and Organisational Performance, 2007.
- [27] A. Felstead, D. Gallie, F. Green, and H. Inanc, "Skills at work in britain: First findings from the skills and employment survey 2012," 2013.
- [28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.
- [29] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.
- [30] A. Joulin, É. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 427–431.
- [31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.
- [32] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.
- [33] M. Li, K. H. Lim, T. Guo, and J. Liu, "A transformer-based framework for poi-level social post geolocation," *Proceedings of the 45th European Conference on Information Retrieval*, 2023.
- [34] L. F. Simanjuntak, R. Mahendra, and E. Yulianti, "We know you are living in bali: Location prediction of twitter users using bert language model," *Big Data and Cognitive Computing*, vol. 6, no. 3, p. 77, 2022.
- [35] J. Arreola, L. Garcia, J. Ramos-Zavaleta, and A. Rodriguez, "An embeddings based recommendation system for mexican tourism," in *IberLEF@SEPLN*, 2021, pp. 110–117.
- [36] N. L. Ho and K. H. Lim, "Poibert: A transformer-based model for the tour recommendation problem," *Proceedings of the 2022 IEEE International Conference on Big Data*, 2022.