2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Matching Recruiters and Jobseekers on Twitter

Aparup Khatua & Wolfgang Nejdl
L3S Research Center, Leibniz Universität Hannover, Hannover, Germany
khatua@L3S.de, nejdl@L3S.de

***Abstract*—An efficient job recommendation framework needs to recommend an appropriate jobseeker to a recruiter and vice-versa. Prior studies have mostly considered datasets from commercial job portals such as LinkedIn or CareerBuilder. However, these datasets are proprietary and not publicly available. Moreover, these portals charge their clients for offering customized services. Hence, we explore whether publicly available Twitter data can be a viable alternative to commercial job portals. We have extracted 0.76 million job-related tweets. We have manually annotated tweet-pairs from recruiters and jobseekers in the domain of computer science jobs. Next, we have employed Siamese architecture and considered multiple artificial neural network models with different word embeddings. We have achieved around 97% accuracy for some of our models. Our study demonstrates the potential of the Twitter platform for job recommendations.**

Keywords — Job recommendation; Twitter platform; Siamese network; Word embedding

I. INTRODUCTION

Recent developments in web technology, social media mining, and artificial intelligence have impacted the job searching process. Commercial job portals have reduced the information gap between recruiters and job seekers. Recruiters heavily use these job portals for their recruitment process. Similarly, job seekers are also exploring these job portals to search for a suitable job. Hence, information science researchers are trying to recommend the appropriate candidate to a recruiter or recommending the appropriate vacancy to a jobseeker. Prior studies have conceptualized this task as a person-job fit problem [3]. Existing literature has mainly considered commercial job portals as their data source. However, job-related data from these portals are proprietary data and not publicly available. Hence, this paper is exploring whether publicly available social media data, such as Twitter, can be a viable substitute for commercial job portals.

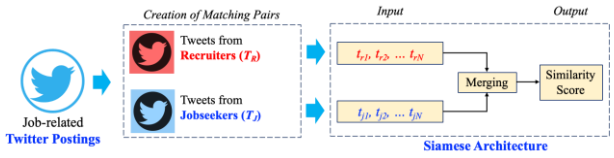

Fig. 1. Our Research Framework

We have formulated our task as a person-job fit problem where we are trying to connect job seekers and recruiters on the Twitter platform. Thus, we have crawled job-related tweets from both recruiters (T_R) and jobseekers (T_J). Next, we need annotated job data for model building. Hence, we have annotated and matched tweet-pairs (t_{r1}:t_{j1}, t_{r2}:t_{j2}, … t_{rN}:t_{jN}) from recruiters and jobseekers. We have used this annotated data for the training and evaluation of our proposed Siamese architecture. Figure 1 graphically represents our overall research framework.

We propose a semi-supervised Siamese architecture-based framework for job recommendation in the domain of computer science-related jobs. We have restricted our analysis to data science, game developers, software engineers, and web developer jobs. However, our proposed framework can be extrapolated to the other domains, such as accounting to advertising jobs or prosecutors to physiologists. On methodological fronts, we have employed Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Bi-directional LSTM (Bi-LSTM), and Bi-LSTM with attention. We have achieved an accuracy of around 97% for Bi-LSTM and Bi-LSTM with attention. The core contribution of our study is to demonstrate that Twitter data can be a viable substitute for commercial job portals in connecting the jobseekers and recruiters. Hence, this study is making a humble contribution in the domain of artificial intelligence for social good.

II. LITERATURE REVIEW

Job-related research is becoming popular in the last few years. This stream of research mostly focused on job recommendation or person-job fit [3]. In addition to person-job fit, prior research in this domain has also probed job mobility [13], salary benchmarking [12], privacy issues [7], personalized question recommendation for an interview [2], and many others. This paper focuses on the person-job fit problem. The efficient job recommendation system needs to connect two different entities: recruiters and jobseekers. Thus, prior research considered two separate sets of data as input variables: one input variable is the candidate profiles, and the other input variable is the job description from the recruiters. However, automated job recommendation is a challenging task. For example, the required skillsets for 'data science' can be very similar to the skillsets for 'data engineering', 'data analysis', or 'machine learning'. A simple rule-based approach might not be the most efficient approach to address this named entity recognition problem. For the sake of brevity, Table I reports the existing literature in tabular format.

Our literature review reveals that prior studies have employed a diverse range of methodology, which ranges from the simple content analysis [8] to CNN-based approaches [3], from statistical relational learning [17] to representation learning [19]. However, prior studies mostly consider datasets that are either proprietary or owned by corporates like CareerBuilder or LinkedIn. Some of these studies are the outcome of their in-house research. Commercial job portals have two significant shortcomings. First, job-related data in these job portals are proprietary; thus, accessing the data through crawling can have legal consequences. Second, most of these portals charge their users. So, big recruiters or high-end jobseekers can afford their customized service, but small organizations and not-so-rich job seekers might not be able to afford it. We probe whether Twitter can address this disparity.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

TABLE I. PRIOR STUDIES ON JOB RECOMMENDATION

Author	Data Source	Methods/Findings
Shalaby et al. (2017)	CareerBuilder	Proposed job recommendation by addressing the short-lived nature of jobs and the rapid rate at which new users and jobs enter the system [20]
Yang et al. (2017)	CareerBuilder	Employed Statistical Relational Learning for developing their job recommendation system and ensured low inappropriate job recommendation [17]
Dave et al. (2018)	CareerBuilder	Proposed a representation learning model which considers information from three networks (job transition network, job-skill network, and skill co-occurrence network) for job recommendation [19]
Geyik et al. (2018)	LinkedIn	Explains the architecture of 'LinkedIn Recruiter product, which enables recruiters to search for relevant candidates and obtain candidate recommendations for their job postings' [15]
Ramanath et al. (2018)	LinkedIn	Considered both deep learning models as well as representation learning approaches for talent search systems at LinkedIn [14]
Zhu et al. (2018)	A Chinese company	Employed CNN for person-job fit. Also, it identifies which all items of the job requirements the person can satisfy [3]
Liu et al. (2019)	CareerBuilder	Proposed a vector representation for both job postings and resumes and considered three information graphs (job-job, skill-skill, job-skill) [9]
Meyer et al. (2019)	Indeed	Content analysis of U.S. healthcare data scientist job postings to understand the job requirements [8]
Ozcaglar et al. (2019)	LinkedIn	Developed an entity-personalized talent search model by combining generalized linear mixed models and gradient boosted decision tree models [1]

To the best of our knowledge, none of the prior studies has explored the Twitter platform for a job recommendation. Hence, our paper attempts to address this research gap by exploring the feasibility of social media data for the person-job fit problem. We are trying to map a tweet from a jobseeker with an appropriate and relevant tweet from a recruiter in the domain of high-end computer science jobs. The following section elaborates on why we have selected Twitter as our data source.

III. DATA: WHY TWITTER?

We find that social media users discuss job-related issues on the Twitter Platform. Most companies, irrespective of their size, are using the Twitter platform for promoting their brand and reaching customers. These companies also use the Twitter platform for sharing vacancies, new jobs, and recruitment plans. Thus, Twitter is becoming a popular communication channel not only for big organizations but also for small organizations, who cannot afford the customized service of commercial job portals to reach the labor market. Moreover, the younger generations are highly active on the Twitter platform. Table II reports some sample job-related tweets and various job attributes in those tweets. Our preliminary analysis of the linguistic contents of job-related tweets reveals that mostly job-related tweets clearly mention the expectations. Many of these tweets mention - What is the overall scope of a particular job? Which location? What skill sets? However, it is essential to note that all job-related tweets are not so informative.

TABLE II. JOB-RELATED TWEETS FROM RECRUITERS

Tweets from Job Recruiters	Job Attributes
With SAS analysis experience, an opportunity to join a lovely agency in SW London in this 9 – 12-month Mat Cover Data Analyst role. #marketingjobs #newjob #dataanalyst #analytics #marketinganalyst #SAS #SASprogramme	Experience, Location, Tenure
Are you interested in designing #fullstack #code for web applications in the financial industry? Do you have experience using #Angular, #typescript, #SQS, #Nodejs or #Oauth? Click here #careers #WisconsinJobs #JavaDevelopment #Engineer #JavaScript #HTML5	Experience, Location, Role
iOS Developer Johannesburg, Gauteng, South Africa We are looking for an iOS developer responsible for the development and maintenance of applications aimed at a range of iOS devices including mobile phones and tablet… #jobs #recruiting #careers	Location, Role, Scope
Charles Taylor PLC are now recruiting Senior #Analyst .NET #Developers to join their business to build new innovative systems for insurance industry and support existing #IT applications. Location London. Competitive salary + package. Apply #jobs	Location, Role, Scope, Salary

Additionally, we find some commercial job portals are very focused. For example, one job portal might be known for only finance-related jobs in one particular location/country. As a result, others might not opt for that particular portal. However, this is not a problem for the Twitter platform. Twitter data comes from all over the world.

TABLE III. SAMPLE TRAINING DATA

Tweets from Recruiter	Tweets from Jobseeker	Label
We're hiring a new web developer! If you or someone you know might be a good fit, take a look at the job posting right here	I'm a graphics designer, database administrator, web developer and designer, android developer. I'm looking to collaborate or work with anyone on any project. Please help	**C**
Cognizant is looking for teammates like you. See our latest #IT job openings, including "Data Analyst", via the link in our bio. #Lynnwood, WA	are you looking web designer and developer? I'm a professional web designer and developer with 2 years of experience working with international clients and agency's.	**W**
I am looking for a fantastic and professional web designer. Focused on online shopping, subscriptions and creativity. PLEASE let me know your favorite recommendations. #webdesigner	I am WordPress web developer I have 3years experience. looking for a working opportunity with agency or team.. #wordpress #hire #WebsiteDesign #webagency #agency	**C**
New job opening! We are looking for a full-time (permanent contract) game designer to join our core team More info on our website, feel free to check it out! #swissgames #gamedev	Hi, I saw that you are looking for expert web designer / developer, Well I can assist you as I have designed / developed 100+ website for various clients with different	**W**

Hence, using Twitter Search API, we have extracted job-related tweets. We have considered a set of crucial keywords for the crawling purpose, such as 'job', 'vacancy', 'hiring', and 'employment'. We have extracted 0.76 million tweets during November and December 2019. This initial corpus has resulted in a diverse set of job-related tweets that range from architectural to accounting jobs.

For the analysis purpose, we have focused on computer science jobs where the authors have the required expertise to correctly annotate and match the expectation of recruiters with the experience of the jobseekers. We have identified four prominent types of computer science jobs: data science, game developer, software engineer, and web developer. We have manually annotated 704 unique tweets (52% of them by jobseekers and 48% of them by recruiters). Next, we have matched a recruiter's tweet with an appropriate jobseeker's tweet and created 3980 recruiter-jobseeker tweet pairs (which includes 38% correctly and 62% wrongly matched pairs) for training purposes. It is worth noting that while every tweet pair is unique, every tweet within the tweet pairs is not. Table III has reported a few correctly (C) and wrongly (W) matched pairs, which we used for the modeling purpose.

IV. METHODOLOGY

Following prior Twitter-based studies, we have preprocessed our corpus. We have followed the standard steps such as tokenization, word normalization, and lowercasing of all words. We have also removed URLs, email-ids, and user handles and replaced these URLs, email-ids, and user handles with blank space. Next, we have employed Siamese architecture, which compares the semantic meaning of two different but similar types of text. In other words, Siamese architecture explores the relation between two different texts based on their semantic meaning [10]. This is commonly known as the text pair comparison. Hence, in our research context, this text pair comparison approach measures the semantic similarity of a tweet pair to determine whether one tweet is closer to another or not i.e. comparing the semantic similarity between the tweets from recruiters with the tweets from jobseekers for developing an efficient job recommendation framework. Existing literature has obtained the state of the art results by using CNN [4] and RNN [5], [11] for the above sentence similarity task. Hence, we have followed a similar approach and considered four models for our analysis: CNN, LSTM, Bi-LSTM, and attention-based Bi-LSTM for text-pair comparison.

Following prior studies, we have implemented the Siamese architecture for our models. We have created two identical sub-networks which read the corpus and generates a fixed representation. In other words, we have considered two identical sub-networks for the tweets from recruiters and jobseekers. Each sub-network reads the tweets and produces its vector representation for the next layer. Both subnetworks share the same weight for comparing a tweet pair (one from the recruiter and the other one from the jobseeker) in the same vector space. We have considered 80% of our corpus for training and the remaining and unexposed 20% of the corpus for testing using stratified sampling for our analysis. As we mentioned earlier, we trained, validated, and tested by using our manually annotated tweet-pairs.

We have considered different pre-trained and publicly available word embeddings i.e., GloVe [6]. In the first step, we represent our tweets in a low-dimensional distributed representation, i.e., word embeddings. Specifically, we use four pre-trained word embeddings as follows: 6B50d (50-dimensional Glove embeddings based on Wikipedia 2014 & Gigaword 5 with 6 billion tokens), 6B100d (100-dimensional Glove embeddings based on Wikipedia 2014 & Gigaword 5 with 6 billion tokens), 27B50d (50-dimensional GloVe embeddings based on the Twitter corpus with 27 billion tokens), 27B100d (100-dimensional GloVe embeddings based on the Twitter corpus with 27 billion tokens) [6]. We have considered these multiple word embeddings to ensure the robustness of our findings.

First, we use a Siamese CNN model to analyze the contextual similarity of our tweet-pairs. The Siamese structure with two identical sub-networks processes the sentences (or tweets) parallelly with identical weights for each layer. Next, our fully connected layers compute the similarity score between two tweets. We have used an Ecludian distance to measure the similarity. We have used the contrastive loss as a loss function and the Adam method to optimize our model's parameters. The last layer in the Siamese architecture decides whether the tweets from the recruiter and the jobseeker are matching correctly or not.

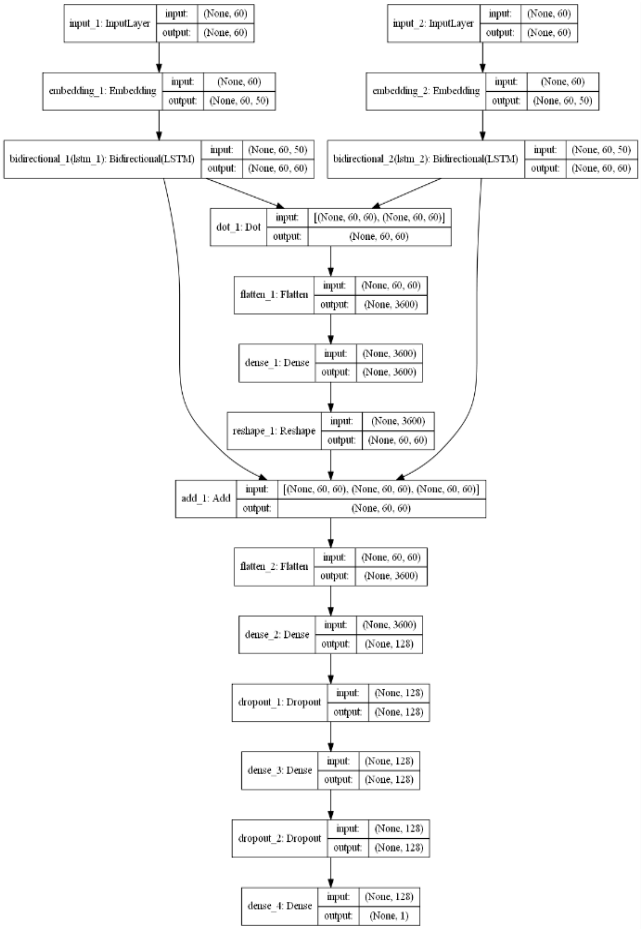

Fig. 2. *Architecture of Siamese Bi-LSTM with Attention*

Second, we consider the LSTM model, which incorporates a gating mechanism to ensure proper gradient propagation through the network [16]. Standard RNN models suffer from vanishing gradient problems and cannot capture the long sequence context, which is a dominant objective for an NLP task. Similar to our earlier model, we have considered the output of two LSTM based sub-network as an input for the next level dense feed-forward layer. This network is comprising of two dense layers with 128 hidden units each. The input strings are also post-padded to produce an equal length sequence. We have considered a dropout rate of 0.45 for our LSTM model. Results remain consistent for other dropout rates in the range of 0.2 to 0.5. Similarly, in addition to 128 hidden units, we have also considered 64 and 256 hidden units, and our results remain broadly consistent.

Third, we employ a Bi-LSTM model to extract the contextual information from both directions. This model's hyperparameters are similar to the previous LSTM model - except the subnetwork designed by the Bi-LSTM unit for improvement. Finally, we incorporate the attention mechanism [18] in our previous Bi-LSTM model to amplify the contribution of critical keywords within a tweet. An attention mechanism assigns a weight to each word, which reflects its importance. This attention mechanism aggregates all the intermediate hidden states using their relative importance and feed-forward to the subsequent dense layer for the classification task. Hence, this approach consists of an embedding layer, a Bi-LSTM layer, an attention layer, and the final classification layer with dropout to prevent the overfitting (refer to Figure 2 for the model details).

V. FINDINGS & CONCLUSION

Table IV reports the classification accuracies. We find advanced sequence-based models have outperformed the CNN model, and higher dimensional word embeddings have outperformed lower-dimensional word embeddings. Our best performing models have reported accuracies around 97%, which is significantly high. Intuitively, tweets are short-texts - in comparison to the long job description and elaborate resumes. Thus, users try to incorporate multiple job attributes within a tweet. Hence, the word overlapping between a correctly matched pair can be potentially high. Therefore, we probe this further.

TABLE IV. CLASSIFICATION ACCURACIES WITH GLOVE EMBEDDING

Model	6B50d	6B100d	27B50d	27B100d
CNN	0.8101	0.8341	0.6202	0.7837
LSTM	0.8832	0.9774	0.6181	0.8065
BI-LSTM	0.9008	0.9422	0.6759	0.9736
Bi-LSTM + Attn.	0.9497	0.9749	0.6093	0.9661

We find that the average word counts of tweets from recruiters and jobseekers are 30.2 and 38.0, respectively. Next, we looked into common words and word share [=common words/(word count of recruiter's tweet + word count of jobseeker's tweet)] between a pair of tweets. We find that the average common words (word share) for correctly matched pairs are 4.39 (6.3%). Similarly, average common words (word share) for wrongly matched pairs are 3.44 (4.8%). Thus, the proportion of common words are not very high in our corpus. Our approach might not be very appropriate for job domains that require soft skills. Soft skills might not be adequately expressed through the Twitter platform. Job-related tweets from different domains and more number of annotated tweets can enhance the robustness of our findings. Future studies also need to consider context-specific word embeddings instead of pre-trained GloVe. These limitations might open up some exciting avenues for future research.

ACKNOWLEDGMENT

Funding for this project was, in part, provided by the European Union's Horizon 2020 research and innovation program under grant agreement No 832921.

REFERENCES

[1] C. Ozcaglar, S. Geyik, B. Schmitz, P. Sharma, A. Shelkovnykov, Y. Ma, and E. Buchanan. "Entity Personalized Talent Search Models with Tree Interaction Features." In *The World Wide Web Conference*, pp. 3116-3122. 2019.

[2] 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*, pp. 2165-2173. 2019.

[3] 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)* 9, no. 3 (2018): 1-17.

[4] H. He, K. Gimpel, and J. Lin. "Multi-perspective sentence similarity modeling with convolutional neural networks." In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1576-1586. 2015.

[5] J. Mueller, and A. Thyagarajan. "Siamese recurrent architectures for learning sentence similarity." In *thirtieth AAAI conference on artificial intelligence*. 2016.

[6] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation." In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543. 2014.

[7] K. Kenthapadi, and T. T. L. Tran. "Pripearl: A framework for privacy-preserving analytics and reporting at linkedin." In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2183-2191. 2018.

[8] M. A. Meyer. "Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings." *Journal of the American Medical Informatics Association* 26, no. 5 (2019): 383-391.

[9] M. Liu, J. Wang, K. Abdelfatah, and M. Korayem. "Tripartite Vector Representations for Better Job Recommendation." *arXiv preprint arXiv:1907.12379*. 2019.

[10] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli. "Semeval-2014 task 1: Evaluation of compositional distributional." *SemEval-2014*. 2014.

[11] P. Neculoiu, M. Versteegh, and M. Rotaru. "Learning text similarity with siamese recurrent networks." In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pp. 148-157. 2016.

[12] Q. Meng, H. Zhu, K. Xiao, and H. Xiong. "Intelligent salary benchmarking for talent recruitment: a holistic matrix factorization approach." In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 337-346. IEEE, 2018.

[13] Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong. "A Hierarchical Career-Path-Aware Neural Network for Job Mobility Prediction." In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 14-24. 2019.

[14] R. Ramanath, H. Inan, G. Polatkan, B. Hu, Q. Guo, C. Ozcaglar, X. Wu, K. Kenthapadi, and S. C. Geyik. "Towards Deep and Representation Learning for Talent Search at LinkedIn." In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 2253-2261. 2018.

[15] S. C. Geyik, Q. Guo, B. Hu, C. Ozcaglar, K. Thakkar, X. Wu, and K. Kenthapadi. "Talent search and recommendation systems at LinkedIn: Practical challenges and lessons learned." In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1353-1354. 2018.

[16] S. Hochreiter, and J. Schmidhuber. "Long short-term memory." *Neural computation* 9, no. 8 (1997): 1735-1780.

[17] S. Yang, M. Korayem, K. AlJadda, T. Grainger, and S. Natarajan. "Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach." *Knowledge-Based Systems* 136 (2017): 37-45.

[18] T. Rocktäschel, E. Grefenstette, K. M. Hermann, T. Kočiský, and P. Blunsom. "Reasoning about entailment with neural attention." *arXiv preprint arXiv:1509.06664*. 2015.

[19] V. S. Dave, B. Zhang, M. A 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*, pp. 1997-2005. 2018.

[20] W. Shalaby, B. AlAila, M. Korayem, L. Pournajaf, K. AlJadda, S. Quinn, and W. Zadrozny. "Help me find a job: A graph-based approach for job recommendation at scale." In *2017 IEEE International Conference on Big Data (Big Data)*, pp. 1544-1553. IEEE, 2017.