Job Advertisements Analysis in Social CRM Industry

Jhanvi Lotwala
Masters student in Data Science
San Jose State University
San Jose, USA
jhanvi.lotwala@sjsu.edu

Dr. Teng Moh
Department of Computer Science
San Jose State University
San Jose, USA
teng.moh@sjsu.edu

***Abstract*—Social CRM or Customer Relationship Management is the incorporation of social media channels into CRM platforms. These platforms help businesses to communicate with the customers using the channel of their choice - phone, text, chat, email or social media. Looking at the huge demand in the online job postings related to Social CRM, analyzing these advertisements provides an idea to the job seekers and hiring team to focus on specific keywords and skills for the job. In this context, this work aims to collect and analyze job advertisements with the help of text-mining techniques. The analysis provided with a systematic approach to understand the skills and knowledge required in this dynamic job market.**

***Index Terms*—advertisements, Social CRM, text-mining**

I. Introduction

There is a significant rise in digital platforms and social media that has changed the way of interaction between companies and the customers. With the increase in the usage of social media websites like Facebook, Twitter, Instagram, and Snapchat every day, businesses have to manage the relationship with the customers in innovative ways. This evolution has led to the necessity in Social Customer Relationship Management (Social CRM) that focuses on engagement with customers directly on social media platforms to understand and satisfy their needs better [1].

Social CRM is about listening to the customer through these online platforms in real-time. It can be through comments, viewership, or shares. This feedback is helpful for the businesses to build stronger relationships with their customers which can improve the future services being offered. A job position related to the Social CRM indicates the job title, job requirements, and job summary overall just like any other job position [1].

To analyze these job positions, we used text-mining which is a technique used to scrape out beneficial patterns and insights from text-based data. It involves a mixture of various fields like data extraction, cleaning and sorting data, machine learning, statistics, and human language to make sense of the text data. All these methods can do a various tasks like summarizing information, categorizing data, grouping of similar items, and more. Text mining focuses on texts that is not organized that often comes from from sources like emails, websites, or social media [2].

The remaining sections of this paper focuses on the Related work, Data Understanding, Baseline Implementation, New Approach, Results, and Conclusion.

II. Related Work

Menezes, et al. [1] focuses on N-grams and topic modeling technique like BERTopic to analyze the job advertisemets in Social CRM industry which also works as a baseline discussed in this research paper. Papoutsoglou, et al. [3] focuses on the Exploratory Factor Analysis (EFA) and Principal Component Analysis (PCA) for analyzing job advertisements in Human Relationship Management industry. The researchers used statistical methods to reduce data complexity and see underlying patterns. By using EFA and PCA, they found out the skills that appear frequently in the job advertisements which suggests the often need. They also did the Kaiser-Meyer-Olkin (KMO) Measure and Bartlett's Test to check if the data were suitable for factor analysis. While Habiba, et al. [4] focuses on the data mining techniques like KNN, Decision Tree, Support Vector Machine, and Random Forest Classifier to predict if the job advertisement was fake or real.

III. Data Understanding

Data were scraped from various job posting websites using Diffbot, a web scraping tool used by the authors of the original research paper [1]. As indicated by the author, the data were almost cleaned and it were made available by the author on the prompt request. So, the data were arranged in a following manner: it has variables like id, requirements, resolvedPageUrl, summary, tasks, text, and title. For the data cleaning, there was a need to remove the duplicates but only based on the variables like id and resolvedPageUrl. For text standardization, all the text was converted to lowercase, and the non-English ads were converted using the googletrans library in Python. There was also a removal of accents, and irrelevant characters from each text column. Whitespaces and alphanumeric characters were kept for data analysis. The common English stopwords and some additional stopwords like “view”, “e g”, “please”, “send”, “end”, “de”, “compliance”, “risk”, “risks”, “use”, “sanctions”, “france”, “united”, “kingdom”, “york”, “philippinesofficer”, “america”, “abbott”, “nbsp”, “ireland”, “northern”, “belfast”, “citi”, “chalhoub”, “bcg”, “per”, “hq”, “rohq”, “fiserv”, “aml”, “surveillance”, “surveillances”, “ct”, “legal”, “regulatory”, “regulations”, “rules”, “law”, “hybrid-belfast”, “citiover”, “tts”, “grct”, “gft”, “email”, “apply”, and “career” also needed to be removed as mentioned by the author [1].

IV. Baseline Implementation

The baseline implementation was followed by the diagram below:

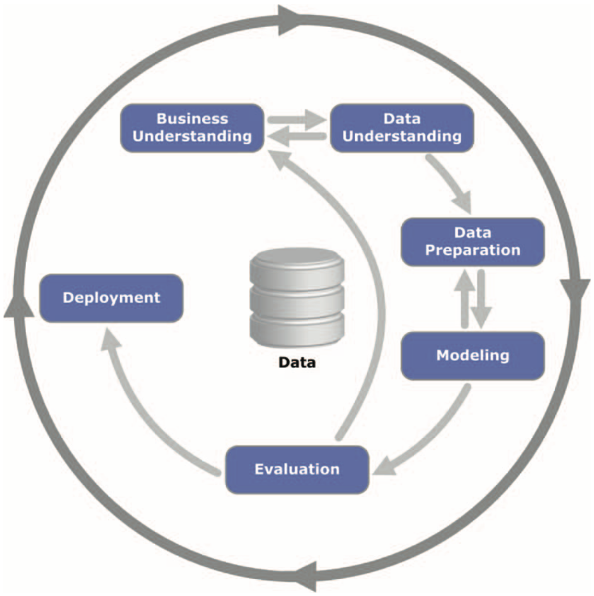

Fig. 1. Overview of the implementation. Adapted from [1]

In order to work only with the text data, id and resolved-PageUrl were removed while keeping the remaining columns. From the remaining columns, the job requirements column indicates the years of experience amongst the text. So, to extract the years of experience, regex (regular expression) was used to only extract the years necessarily digits while leaving rest of the text alone. After extracting the years, the values corresponding to each digit were counted. There were some outliers like some unusual numbers so those were included in the 10+ years of experience bucket. To see the frequency of each year that was extracted from the job requirements, please refer the Fig. 2. Then, we used the word cloud by utilizing the WordCloud library in Python. It highlights the most frequent words seen in the text. So, the word cloud of job requirements can be seen in Fig. 3.

The next task was to see the most frequent job titles in the data. So, we worked on the generation of n-grams to visualize the most frequent words for the job title column. In order to generate n-grams, tokenization and some text preprocessing was needed. We put the entire job title column into a corpus and then tokenization of words was to be done. Tokenization necessarily divides the sentences into bag of words or you can say an array of all the words in the corpus. Now, as we had the bag of words representation, it was easier to generate n-grams where we used the bi-gram which means separating the words into set of 2. Then by counting the frequency of words in the n-grams, we were able to find out the job titles that were essential in the Social CRM industry. Also, in the text preprocessing we had to remove some words that mean the same comparing to other words in the text. For example, ’manag’ and ’manager’ or ’mgmt’ and ’manager’. Please refer the Fig.4 to see the n-grams based on the job titles.

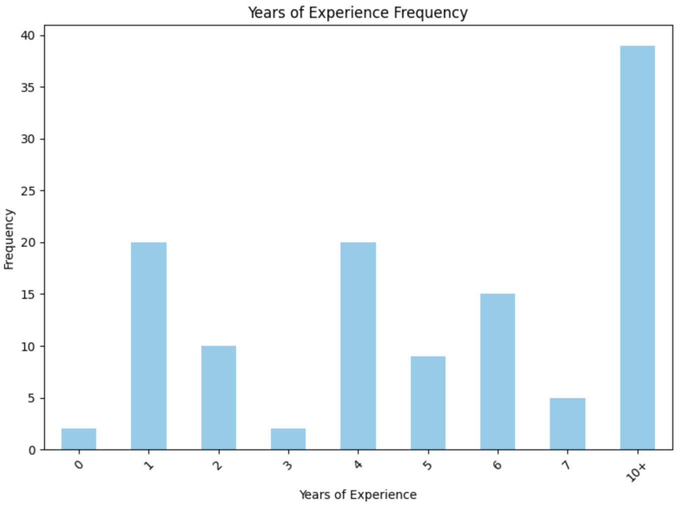

Fig. 2. Years of Experience.

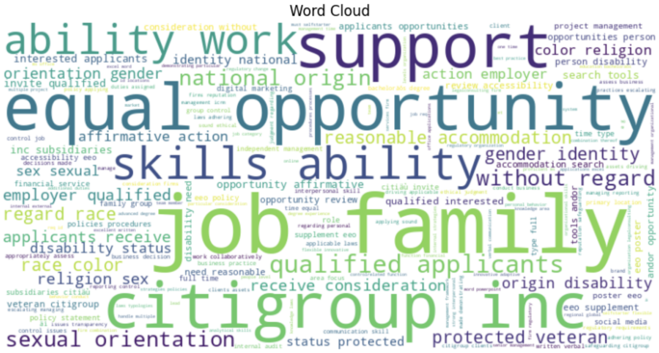

Fig. 3. Word Cloud of job description.

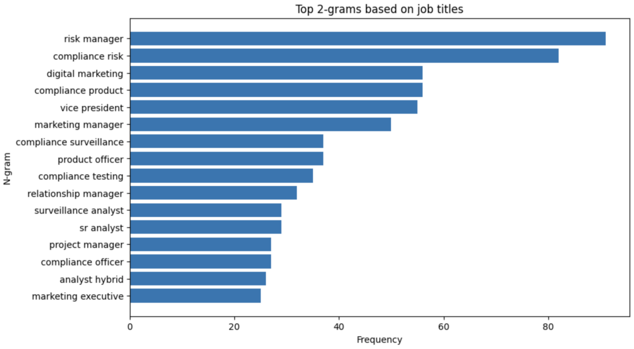

Fig. 4. Word Cloud of job requirements.

In the baseline, the authors have used the BERTopic Modeling technique which is an advanced technique in the field of Natural language Processing. BERTopic model is helpful in visualizing the meaning-based relationships between the words or to see the co-occurrence between the words similar to how we see the correlation between the numbers. BERTopic applies clustering algorithms that can help visualize different clusters of topics. Here, we used the job summary column as it summarizes the overall job description to visualize the most important topics or essentially keywords in the job description. Please refer the Fig.5 and Fig.6 to see the most important keywords in the job advertisements for Social CRM industry.

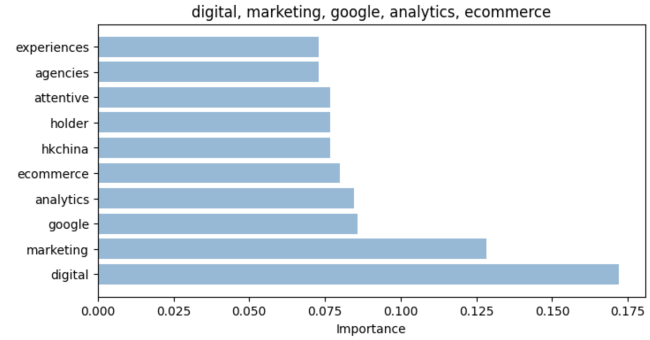

Fig. 5. Topic Frequency.

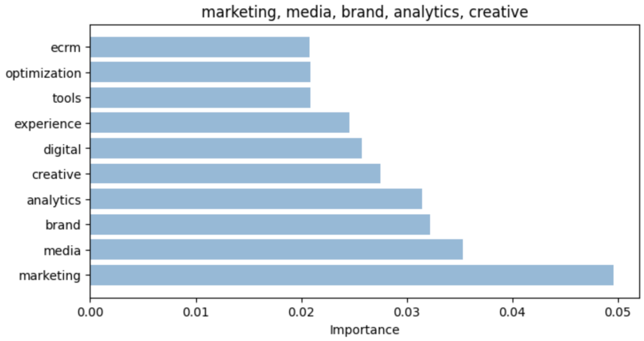

Fig. 6. Topic Frequency.

V. New Approach

For visualizing the separation between certain words or topics that might be in a particular group in the data, Clustering along with Dimensionality Reduction techniques like DBSCAN clustering with PCA and t-SNE have been used. PCA focuses on extracting the first few principal components related to the most important variables in the data. Here, we consider the n-grams that we generated earlier based on the job titles to form the clusters. The visualization in Fig.7 provide insights into the underlying structure of the job title n-grams. Clusters may represent specific industry sectors, job functions, or common phrases in job titles. The dissimilarity between groups of n-grams is indicated by the separation between the clusters. Good separation would generally suggest that the clustering process has been effective in differentiating between different topics in the data. PCA usually see the linear patterns in the data while t-Distributed Stochastic Neighbor Embedding (t-SNE) focuses on finding the non-linear patterns in the data.

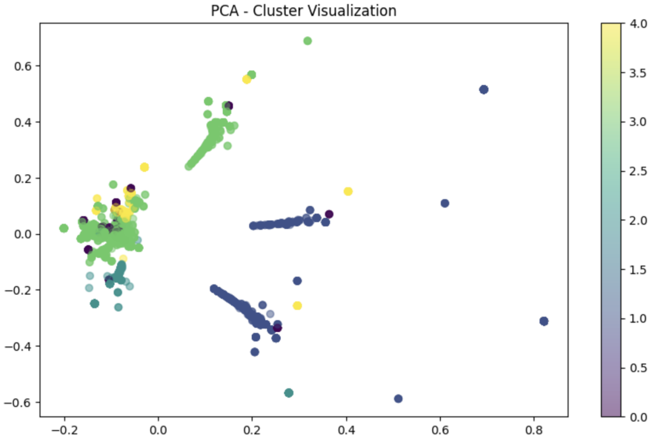

Fig. 7. PCA cluster.

As our data consists of a lot of noise, it seemed better to use non-linear dimensionality reduction technique like t-SNE. Before using t-SNE, we utilized the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm which do not require specifying the number of clusters and is generally good at handling noise unlike K-Means or Spectral Clustering algorithms. After using any clustering algorithm, we calculate the silhouette score which measures how similar an object is to its own cluster when compared to other clusters. It usually ranges from [-1,1]. +1 would indicate that clusters generated are effective although -1 would indicate that clustering is not being done correctly. For DBSCAN, the silhouette score came out to be 0.3043 which is much better than what we experimented with K-Means and Spectral Clustering.

In order to use DBSCAN with t-SNE, first t-SNE is applied to the high-dimensional data to decrease it to fewer dimensions while preserving the local structure i.e. keep the neighbors together. After the data is reduced, DBSCAN is applied to the output of t-SNE. Since t-SNE preserves local structure, it enhances the performance of density-based clustering algorithms like DBSCAN, that aids in identifying clusters in the transformed space. In the visualization of Fig.8, each color represents a different cluster assigned by DBSCAN. There are several distinct clusters with many points close together suggesting that there are well-defined groupings in the data. This aligns with the decent silhouette score of 0.3043, indicating a moderate separation within clusters.

VI. Results and Conclusion

The results will focus on the overall data analysis. In the beginning, we see that the years of experience summary helps

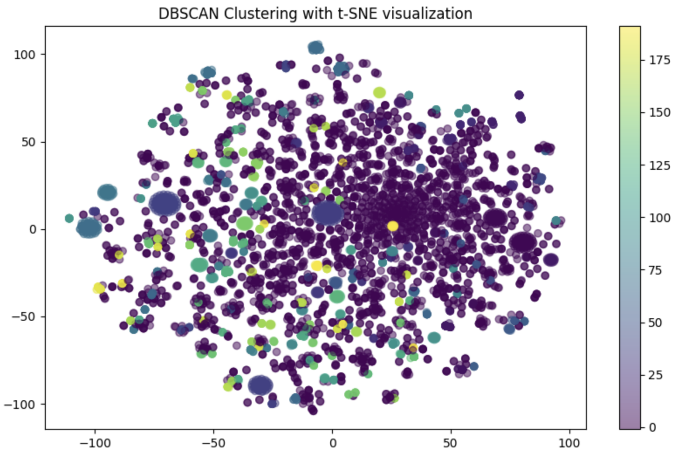

Fig. 8. DBSCAN cluster with t-SNE.

us visualize that there were many job postings that required experience of more than 10 years which suggests that the data is skewed towards the higher years of experience. The word cloud indicated the highlights in the job description so the important keywords seem to be ’equal’, ’opportunity’, ’skills’, and more. This can be considered an important factor for the job seekers. In the N-gram representation of job titles, ’risk manager’, ’compliancy risk’, and ’digital marketing seem to be the highest keywords in the job title. It suggests that the job seekers may focus on the skills related to these positions. Then BERTopic modeling gave an importance score to the keywords that were most important in the data. It turned out that ’digital’, ’marketing’, ’google’, and ’analytics’ seem to be the popular choices based on the importance scores. It gives a hint to the hiring team to look for such keywords in the resumes of the job seekers. For the new approach, PCA was used with K-Means clustering which helped us see some reasonable patterns in the data. The experiments were done with clustering so the silhouette scores were compared between K-Means, Spectral Clustering, and DBSCAN. And the highest silhouette score of 0.3043 was achieved by DBSCAN compared to silhouette scores of K-Means and Spectral Clustering that were close to -1. Therefore, t-SNE method was used with DBSCAN clustering and we were able to see some underlying structure in the data. Although, it would require some more techniques to visualize better. Although, the techniques used in this research could have helped more if there was an exact detection of noise because the authors of the original research paper spent a period of more than a year to extract the data and clear the noise.

Acknowledgment

I would like to acknowledge the authors of the original research paper, especially Fabio M. F. Lobato to keep in constant touch whenever needed. And I would like to acknowledge Dr. Teng Moh for his guidance throughout the semester in the course of CS 274 at San Jose State University.

References

[1] P. C. Menezes, A. F. Justino, B. A. Barata, A. F. Jacob Junior and F. M. Lobato, ”Market Overview in Social CRM: An Analysis of Job Advertisements,” in 2023 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Venice, Italy, 2023, pp. 555-562, doi: 10.1109/WI-IAT59888.2023.00092. keywords: Industries;Social networking (online);Taxonomy;Customer relationship management;Companies;Market research;Intelligent agents;Data Analysis;Social CRM;Job Analysis;Text Mining;Market Analysis

[2] Preeti, ”Review on Text Mining: Techniques, Applications and Issues,” 2021 10th International Conference on System Modeling and Advancement in Research Trends (SMART), MORADABAD, India, 2021, pp. 474-478, doi: 10.1109/SMART52563.2021.9676285. keywords: Text mining;Social networking (online);Decision making;Web mining;Data collection;Market research;Feature extraction;Text Mining Process;Techniques;Summarization;Applications

[3] M. Papoutsoglou, N. Mittas and L. Angelis, ”Mining People Analytics from StackOverflow Job Advertisements,” 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Vienna, Austria, 2017, pp. 108-115, doi: 10.1109/SEAA.2017.50. keywords: Companies;Data mining;Market research;Data analysis;LinkedIn;Recruitment;Software;people analytics;job advertisements;professional social networks;e-recruitment;competence mining

[4] S. U. Habiba, M. K. Islam and F. Tasnim, ”A Comparative Study on Fake Job Post Prediction Using Different Data mining Techniques,” 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh, 2021, pp. 543-546, doi: 10.1109/ICREST51555.2021.9331230. keywords: Support vector machines;Deep learning;Machine learning algorithms;Neural networks;Data mining;Task analysis;Random forests;false job prediction;deep learning;data mining