A Novel Approach for Job Mining and Trend Summarization on Social Media Posts

Ramdas Gawande, Nilesh J. Uke

Department of Information Technology,
P. C. C. O. E., Pune, India
Savitribai Phule Pune University, Maharashtra, India
ramdas.gawande@gmail.com, nilesh.ukel@gmail.com

Abstract—In the job classification field, precise classification of jobs to profession categories is important for harmonizing job seekers with appropriate jobs. An example of such a job title classification system is an automatic text job post classification system that utilizes machine learning. Machine learning based job type classification techniques for text and related entities have been well researched in academia and also been successfully applied in many industrial settings. Digital recruitment is a popular online method that has been widely used for attracting individuals who are seeking for career opportunities. In recent years digital recruitment is transforming from passive websites such as Monster and Career Builder.

In this paper we present a novel approach, a machine learning-based semi-supervised job title classification system. Our method leverages a varied collection of classification and techniques to tackle the challenges of designing a scalable classification system for a large taxonomy of job categories. It encompasses these techniques in cascade classification architecture. We first present the architecture of our system, which consists of a two-stage Capture with filtration and fine level classification algorithm. The paper concludes by presenting experimental results on real world live data.

Index Terms—Big Data, Cloud Computing, Hadoop, HDFS, Data Analysis and Machine Learning.

I. INTRODUCTION

Social media has become very popular communication tool among Internet users in recent years. A large unstructured data is available for analysis on the social web. The data available on these sites have redundancies as users are free to enter the data according to their knowledge and interest [1]. As social media has exploded in popularity, efforts are being made to use social media to recruit for new career opportunities, particularly LinkedIn and Twitter have enabled an emerging trend for matching individuals with possible opportunities based on their interests. This makes finding more relevant jobs easier for both employees and employers [2]. The rapid growth of social networks in recent years has developed a new business: the trade of social networks users data. These social networks data are becoming important for many companies around the world and are often used to determine social networks users interests for items in order to propose or advertise items to them [3]. Social network sites are web-based services that allow users to construct profile (public or semi-public), to share connections with other users and Rick to view and traverse lists of connections made by others in the system. The personal information posted by users of a social network (which may involve personal description, posts, ratings, but also social links) can be exploited by a recommender system [3, 4].

There are various passive job portals sites are available in the market. Now a days jobs seekers are moving towards more of an active social sites e.g. LinkedIn, Facebook etc for taking various job opportunities. A job recommender system is software that elicits the interests or preferences of individual job seekers for various technological categories, either explicitly or implicitly, and makes recommendations accordingly [5, 6]. Recommender systems are mainly related to information retrieval, machine learning and data mining.

II. PROPOSED SYSTEM

This work proposes an efficient way to process the unstructured job related real-time data, collected from twitter for extracting the knowledge and finding out job pattern/trend analysis.

- Pre-processing of raw and real-time job related data from Social Networking Site.
- To apply the NLP APIs for Text Classification
- To Extract the Knowledge from this processed Data by merging the multiple deep learning API’s for Text Classification
- To design text classifier for job classification
- To extract Job patterns related to Technology and Science

Our Approach Follows,
- To process the live remote feed to prevent unwarranted data loss.
- Analysing the data to make decisions based on real-time processing

A. Real Time Data

- Live Feed From Social Network.
- Data Collected on the Basis of Hashtag.

B. Problems to Process Real Time Data

- Multiple Languages in the Data feed.
- Uneven Structure of the Data.
- High Velocity of Data.
C. Cloud to Process Real Time Data

- To store large amount of data in cloud for Further Processing.
- Cloud helps in Managing this data for Process Scheduling.

Figure 1 shows the actual flow and modules involved in the overall architecture.

![Diagram](image.png)

**Fig. 1. Proposed System Architecture**

III. IMPLEMENTATION

This paper implements a probabilistic based and keyword-based recommender systems using Twitter-based live data model. Public streaming API of Twitter is used to fetch the live tweets. These tweets are in unstructured format and are converted into appropriate key value pair and stored into Mongo DataStore. Only the job-related tweets are considered and filtered out which are necessary for our work. The job postings are filtered based upon the technology and science category since we are only interested in technology-related jobs. These job postings are classified into various job categories and technologies for example Java, python, android, C++, iOS. The system tries to understand the context of the sentence or post and classify it accordingly. It uses various natural language processing approaches (streaming, lemmatisation, dictionary lookup) to improve the classification. Different deep learning text classification APIs are used and compared to find out performance metrics with respect to classification accuracy and time needed for processing per thousand job postings. Meaning cloud text classification method is compared with rosette text classification API. Both the approaches of classification process the job-related tweets and find the relevance with respect to science and technology category. Following graph shows the experimental results which can help us to understand the current job trends in the market.

**Algorithm: Filtration.**

**Input:** Live Data Feed.

**Steps:**
1. Filter related data.
   1.1 Remove URL.
   1.2 Remove Special Characters.
   1.3 Emotions and smiles.
   1.4 Re-tweets analysis.
2. Divide the Data into Appropriate Key Value Pair.

**Output:** Filtered data.

**Algorithm: Analysis and Classification.**

**Input:** Filtered Data.

**Steps:**
1. Gather the filtered data from Data Store.
2. Apply NLP using Machine Learning APIs for Individual Data Item from Data Store.
   2.1 Meaning Cloud.
   2.2 Rosette.
3. Persist the final summary into data store.

**Output:** Analyzed and Classified Data.

**Algorithm: Job Trend Summarization.**

**Input:** Analyzed and Classified Data.

**Steps:**
1. For each job event data or for the Technology data, Technology wise Categorical Data is extracted.
2. Summarize the data for all the live feed.
3. Persist the data into data store.

**Output:** Trend Summarization for each job Category.

IV. RESULT

Below pie chart shows the technological trend summery of the tweets taken for a specific period.

Below two graphs shows the comparison between two Machine learning APIs with respect to Processing Time taken and number of relevant jobs processed respectively.

V. CONCLUSION

The proposed system will provide secure and We have developed a job classification system which works on the job posting data from twitter. The various natural language processing tasks are applied on the job posting information retrieved from twitter and deep machine learning APIs Meaning-cloud and Rosette are used for job classification. Our work also produces a trend summery of various jobs under technology and science category. This will help the job seekers, recruiters and educational institutes to understand the current job trend in the market and plan their activities accordingly. We found
that our approach of using NLP and Deep Machine Learning APIs helps the digital recruitment process by processing the unstructured data from twitter. Below is a summary of the work we had done.

- The unstructured data is analyzed and classified to job search entity.
- Job Trend Analysis can be done for various Job categories on Social Media Data feed.
- Various deep machine learning API techniques helps to improve the performance in job classification domain.
- Innovative approach of transforming from passive job websites such as Monster, CareerBuilder to an active form such as Twitter, LinkedIn, Facebook can help the digital recruitment.

REFERENCES

[1] Puneet Garg, Rinkle Rani, Sumit Miglani, "Mining Professional's Data from LinkedIn", Fifth International Conference on Advances in Computing and Communications, 2015.

