DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4213

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# CLASSIFICATION OF JOBS USING LIVE DATA IN CLOUD COMPUTING

Falak Khursheed
CSE DEPT. Integral University
Lucknow, India

Mohd.Shahid Hussain
CSE DEPT. Integral University
Lucknow, India

***Abstract:*** Social media has become very popular communication tool among internet users in the 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. This data needs to be normalized before doing any analysis due to the presence of various redundancies in it as assets of real time digital world daily generate massive volume of real-time data. In the job classification field [1], accurate classification of jobs to line of work categories is important for matching job seekers with appropriate jobs. An instance of such a job title analysis system is a computerized text job post distribution system that uses machine learning. Machine learning based job distribution techniques for text and associated entity have been well studied in the academic world and have also been strongly applied in many industrial environments. In this paper, we introduce a new procedure, machine learning-based semi-supervised job title distribution system [3]. Our system influences a diverse collection of distribution and procedures to deal with the difficulties of designing a scalable distribution system for a large classification of job categories. It incorporates these techniques in cascade classification structure. We first present the structure of our system, which consists of a two-stage Acquisition with filtration and fine level classification algorithm. The paper concludes by presenting preliminary results on real world live data.

***Keywords:*** Big Data, Cloud Computing, Data analysis and Machine Learning.

## I. INTRODUCTION

The improved use of social channels in recent years has uncovered a new business: the enterprise of social networks user's data. These social data are becoming essential for many corporations as well as the industry throughout the world for the scope of employment and are often used to find out the importance of user's for items in order to recommend or advertise items to them. Social network sites are web-based assistance that allows users to build the profile (public or semi-public), to share links with other users and to view and navigate lists of contacts made by others in the system. The private message posted by users of a social network (which may involve personal description, posts, ratings, but also social links) can be used by a recommender system [2]. There are various passive job portals sites available in the market. Now a day's jobs seekers are approaching towards more of an active social sites e.g. Twitter, LinkedIn, Facebook etc for taking various job opportunities. Twitter is one of the common online social network services that provide the facility of communication. It enables users to read and send message of length 140 character. There are about 500 millions of users.In this paper we have presented the idea to process online jobs using live data from twitter. A job recommender system [3], is software that elicit the interests or preferences of individual job seekers for various technological categories, either explicitly or implicitly, and makes recommendations accordingly . Recommender systems are mainly related to information retrieval, machine learning and data mining.

## II. RELATED WORK

Puneet Garg, Rinkle Rani, Sumit Miglani, "Mining Professional's Data from LinkedIn" have used LinkedIn Api for data collection [3]. All the data's are normalized by removing redundancies and LinkedIn connections using geo coordinates. Those data are clustered using hierarchical and K means clustering according to job title and company name. Likewise in many research work social platforms like Twitter, Face book and LinkedIn are used for digital recruitment that has chosen to formalize the textual content of job and provide relevant information expressed by job seekers to provide the job related to users profile [4, 5, 6] . Not only in digital recruitment has social media platform been used in research areas like Transportation. A survey has been performed to check the travel behavior of people all around such as trip purpose, mode of transport, destination choice etc. and as a result the survey has revealed positive view of such data source [7] .This type of research work has used social media in almost every small to bigger aspects of social life. Whether it is about collecting real time data from social websites for flood information [14, 15, 16], to check the details about the damage and hazards in particular area or to use data from these platforms to check the most common prevailing health problem all around [8, 9,11,12,13].

## III. PROPOSED WORK

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. The work focuses on,

- Pre-processing of raw and real-time job related data from Social Networking Site.
- To apply the NLP API's 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 Strategy Follows,

• To process the live remote feed to prevent undesired data loss.

• Study of data to make decisions based on real-time processing

1. Real Time Data:
   - Live data From Social Network.
   - Data Collected on the Basis of # Hashtag.
2. Challenges with Process Real Time Data:
   - Multiple Languages in the Data feed.
   - Uneven Structure of the Data.
   - High Velocity of Data.
3. Cloud to Process Real Time Data:
   - To store a large amount of data in the cloud for More Processing.
   - Cloud helps in Maintaining this data for Process Scheduling.

**Block Diagram**

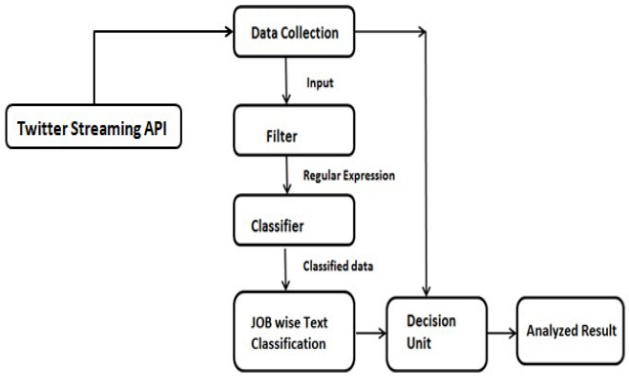



Figure 1 – Block Diagram

## IV. ALGORITHM

**Algorithm : Filtration** [10]

*1) Input: Live Data Feed*
*2) Steps:*
*3) 1. Filter related data*
*4) Remove URL*
*5) Remove Special Characters*
*6) Emotions and smiles*
*7) Re-tweets analysis*
*8)* Divide the Data into Appropriate Key Value Pair.\

Output: Filtered data

**Algorithm : Analysis and Classification** [10]

*1) Input: Filtered Data.*
*2)Gather the filtered data from Data Store*
*3). Apply NLP using Machine Learning API's for Individual Data Item from Data Store*
*4) Meaning Cloud*
*5) Rosette*
*6) Persist the final summary into data store*

Output: Analyzed and Classified Data

**Algorithm: Job Trend Summarization**

*1) Input: Analyzed and Classified Data*

*2) For each job event data or for the Technology data, Technology wise Categorical Data is extracted.*
*3) Keyword based search (Java,Python,android,C++,iOS)*
*4) Summarize the data for all the live feed.*
*5) Persist the data into data store.*

Output: Trend Summarization for each job Category.

## V. FLOWCHART

In suggested system for examining real time as well as offline job-related data for real-time applications using Big Data we have distributed real time Big Data processing design[3] into three parts, i.e.,

1) Data Acquisition Unit

2) Data Processing Unit and

3) Data Analysis and Decision Unit.

In these three units, different algorithms or methods will be mentioned on data for its analysis.

### Data Acquisition Unit

The need for identical processing of the large volume of data was required, which could efficiently examine the Big Data. For that reason, the suggested unit is introduced in the real time Big Data processing structure that collects the massive volume of data from different available data collection unit around the system.

### Data Processing Unit

A data processing unit has two basic functionalities filtration and load balancer. Filtration largely comprises filtration of data and load balancing of processing command. Filtration initiates a process of filtering data which is helpful for analysis and prevents other unrelated data. It has handled to improve a performance of a system as we are simply dealing with the user data.

**Data Analysis and Decision**

This unit comprises three major functions, such as collection and compilation server, results from the storage server, and decision-making server.

When results are to be given to the compilation the data is not in aggregated mode. It is essential to make the given data in aggregated mode for proper storage and processing.

**Flow chart and Activity Diagram**

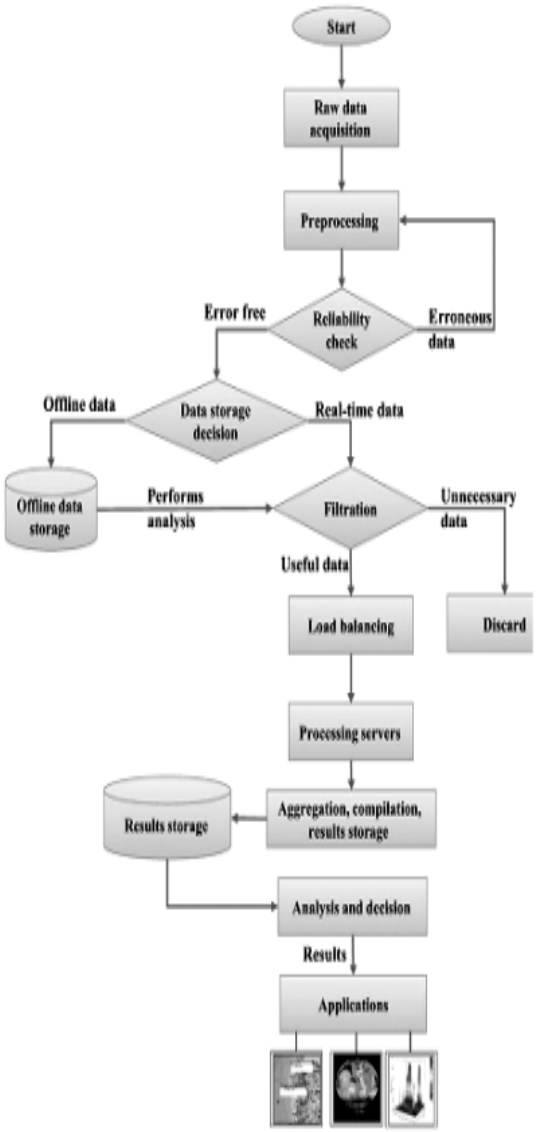



Figure 2 : Flow Chart

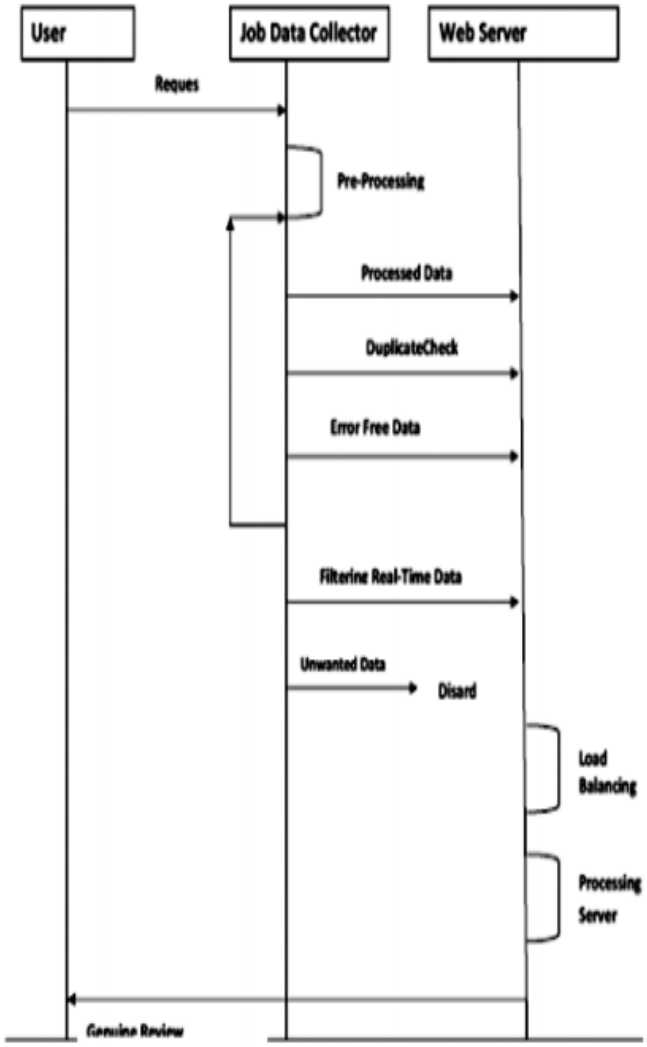


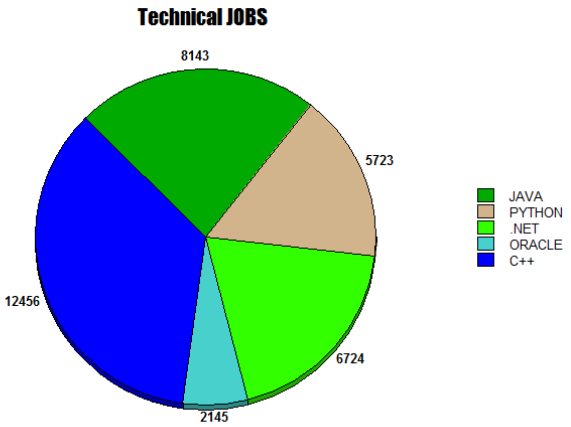

Figure 3: Activity Diagram

**VI. IMPLEMENTATION**

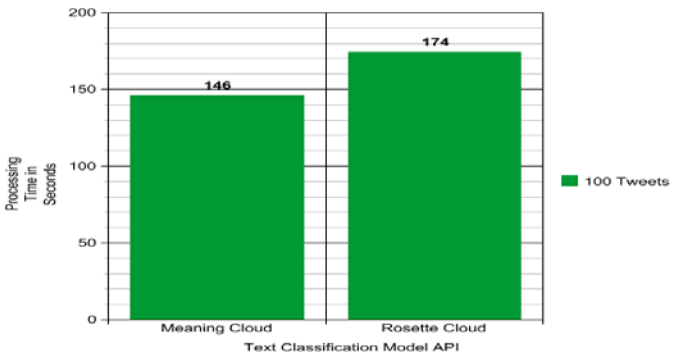
This paper implements a probabilistic based and keyword-based recommender systems using Twitter-based live data model. Public streaming API of Twitter is used to fetch the live tweets. These tweets are in unstructured format and are converted into appropriate key value pair and stored into Mongo Data Store. 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, lemmatization, 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.

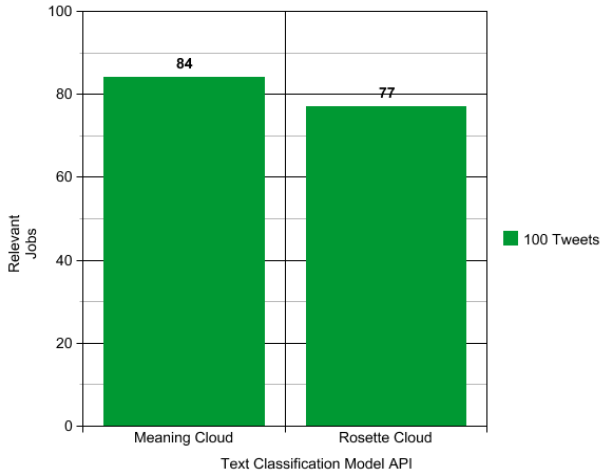
Meaning cloud text classification method is compared with rosette text classification API. Both the approaches of classification process the job-related tweets and find the relevance with respect to science and technology category.
Following graph shows the experimental results which can help us to understand the current job trends in the market.




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




Graph 1: comparison with respect to processing time taken




Graph 2: comparison with respect to number of relevant job

## VII. CONCLUSION

This work offers the probabilistic method to detect job Postings from the live data feed. We originate an aggregated function mining methods for jobs distribution according to the name that they describe pattern behaviors, so as to evaluate our proposed method that escorts user evaluation on a live data set containing reviews of various types of jobs. We found that proposed methods generally exceed the baseline method based votes. As part of future work, we can organize job feed detection into the several other useful job aggregators and vice versa. Exploring ways to learn operation patterns related to that mining so as to advance the accuracy of the current regression model is also an exciting research direction.

## VIII. ACKNOWLEDGEMENT

There are many people, I would like to thank for their help and support in writing this thesis. First of all, I would like to express special thanks to almighty "GOD" for his blessings. I would like to thanks my loving parents and Husband for their love, care, constant encouragement, blessings and cooperation given by them. I express the deep sense of gratitude to all my family members.
I feel extremely privileged to express my deep sense of gratitude and regard to my supervisor **Mohd.Shahid Hussain,**Assistant professor ,Integral university, Lucknow and **M.Haroon** Associate Professor, Department of Computer science & Engineering, Integral university, for their expert supervision, valuable and critical suggestions. I will always remain indebted to them for their undivided attention pointed thrust and constant encouragement in my current endeavor.
At last I would like to thank all those who could not find a separate mention but have helped me directly or indirectly.

## IX. REFERENCE

[1] Javed, Faizan, Qinlong Luo, Matt McNair, Ferosh Jacob, Meng Zhao, and Tae Seung Kang. "Carotene: A Job Title Classification System for the Online Recruitment Domain", 2015 IEEE First International Conference on Big Data Computing Service and Applications, 2015.
[2] Namrata Gawande, Ramdas Gawande, "Processing of Real Time Big data for Machine Learning", May 2016 International Journal of Advanced Research in Computer and Communication Engineering.
[3] PuneetGarg, Rinkle Rani, SumitMiglani, "Mining Professional's Data from LinkedIn", 2015 Fifth International Conference on Advances in Computing and Communications.
[4] Ahmed AbdeenHamed, Xindong Wu, James R Fingar., "A Twitter-based Smoking Cessation Recruitment System",2013

IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

[5] MamadouDiaby , Emmanuel Viennet, "Taxonomy-based Job Recommender Systems On Facebook and LinkedIn Profiles", 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS).

[6] Emmanuel Malherbe, MamadouDiaby , Mario Cataldi , Emmanuel Viennet , Marie- AudeAufaure, "Field Selection for Job Categorization and Recommendation to Social Network Users", 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014).

[7] Exploring the capacity of social media data for modeling travel behavior: Opportunities and challenges. Taha H. Rashidi a,, Alireza Abbasi b, Mojtaba Maghrebi c, Samiul Hasan d, Travis S. Waller.

A.School of Civil and Environmental Engineering, UNSW, Australia. B. School of Engineering and InformationTechnology,UNSW,Australia.

C.Department of Civil, Environment, and Construction Engineering, University of Central Florida, United States.D.Department of Civil Engineering, Ferdowsi University of Mashhad, Mashhad, Khorasan Razavi, Iran.

[8] Using Social Media and Satellite Data for Damage Assessment in Urban Areas During Emergencies.
Guido Cervone, Emily Schnebele, Nigel Waters, Martina Moccaldi, and Rosa Sicignano.

[9] Using linguistic and topic analysis to classify sub-groups of online depression communities.
Thin Nguyen , Bridianne O'Dea , Mark Larsen , Dinh Phung , Svetha Venkatesh , Helen Christensen.

[10] Rathore, Muhammad Mazhar Ullah, Anand Paul, Awais Ahmad, Bo-Wei Chen, Bormin Huang, and Wen Ji. "Real-Time Big Data Analytical Architecture for Remote Sensing Application", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015.

.[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[12] Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In Proceedings of International AAAI Conference on Weblogs and Social Media, 2014.

[13] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In Proceedings of Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015

[14] Tate C, Frazier T (2013) A GIS methodology to assess exposure of coastal infrastructure to storm surge & sea-level rise: a case study of Sarasota County, Florida. J Geogr Nat Disasters1:2167–0587.

[15] Cutter SL (2003) Giscience, disasters, and emergency management. Trans GIS 7(4):439–446.

[16] Dashti S, Palen L, Heris M, Anderson K, Anderson S, Anderson T (2014) Supporting disaster reconnaissance with social media data: a design-oriented case study of the 2013 Colorado floods. In: Proceedings of the 11th international conference on information systems for crisis response and management. ISCRAM, pp 630–639.