ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Behavioral Analysis of Candidates using Sentiment Analysis and Emotion Mining for Recruitment

Sagar S. Patil[1], Pravin S. Game[2]

M.E. Student, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India[1]

Assistant Professor, Dept. of Computer Engineering, Pune Institute of Computer Technology, Pune, India[2]

ABSTRACT: Today's organizations are assessing candidates based on inadequate criteria to determine candidate's behavior for recruitment. We have proposed an iterative behavioral model that assesses candidates by analyzing their tweets. It identifies emotions and polarity from affective text, classifying candidates to proper emotion and polarity classes. Manually set emotion labels using hash tags by candidates forms only a tiny dataset and partially meaningful. But, extracting sentiments and emotions from entire tweets is more meaningful. Hence, generated model is precise and used in making predictions about candidate's behavior. The proposed system uses memory speed virtual distributed storage system to store data in-memory. It uses the concept of the lineage that reduces writes workloads and guarantees timely recovery. Candidate's tweets are stored in Alluxio and processed distributively by Apache Spark. Classifiers are built to operate distributively reducing overall training time. They are compared based on performance criteria. It is observed that Multinomial Naive Bayes and Logistic Regression performs with higher accuracy and less false positive rate for emotion and polarity dataset respectively. It is also important to secure sensitive tweets of candidates. Therefore, Alluxio cluster is kerberized to establish secure communication between the storage system and computational framework. Predictive analytics on time series emotion and polarity data predict candidate's future behavior.

KEYWORDS: Emotion Mining, Kerberos, Sentiment Analysis, Big Data Analytics, Apache Hadoop, Alluxio, Apache Spark, Predictive Analytics.

I. INTRODUCTION

Today's organizations are assessing their candidates they want to recruit and employees based on their compatibility to organization's culture and fitness in a role. They use multiple tests that consist of several roles scenarios as well as capability statements for assessment. Assessing candidate's behavior is equally important to determine mindset and thoughts towards recent and trending topics. This can be done in two ways: First, behavioral assessment test that comprises of some questions that is used to observe, describe candidate's behavior. Second, building an iterative behavioral profile by analyzing candidate's profile on social media sites, their reaction to recent trending topics. Reactions can be mapped to emotion and polarity classes, which then give a more precise in-depth understanding of candidate's behavior.

Assessment results along with the inclusion of behavioral profiling of candidates is recorded by HR. This helps organizations in forming proper structure, allocation of candidates to proper product development and operational efficiency. Predictive analytics combined with proposed behavioral model gives organizations essential information regarding internal candidates and helps in making key decisions like customized employment value proposition. Tweets of candidates are collected, that becomes our input dataset for a particular candidate. Dataset persists in HDFS, thus provides information availability even if one of the connected node goes down. Only selected data from the dataset that needs to be analyzed is brought down to Alluxio, virtual distributed storage system [11].

Alluxio storage system stores in-memory data, so only limited data can be stored in the memory layer of the storage system. Therefore, Alluxio makes use of optimized allocators and evictors to move data between memory and underFS

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

storage layer like HDFS. A candidate is classified to emotion and polarity classes based on trained model with higher accuracy. We can clearly see the difference in the behavior of candidates as we are analyzing social media profile year-wise, month-wise and day-wise. It helps organizations to see patterns in the behavior of candidates.

The behavioral model uses Multinomial Naive Bayes for emotional classification and Multinomial Logistic Regression for polarity classification. Classifier executes as a Spark job which runs distributively on several connected commodity hardware configured nodes. To predict future behavior of candidate, Autoregressive Integrated Moving Average model is used. It operates on time series data of emotion and polarity categories collected over several years. Spark does not provide any authentication among its components while running job. Therefore, it is necessary to provide mutual authentication among server nodes, Hadoop components – Domain specific as well as infrastructure components. Mutual authentication is achieved by integrating Kerberos [13] into Hadoop, Spark, and Alluxio.

II. RELATED WORK

A. Authentication methods for Hadoop

Multiple clients submit their MapReduce jobs for processing. Before submitting any jobs, a client needs to get authenticated by Authentication Server. Somu et al [5] provide a method which is symmetric key based. It uses single authentication factor, supports only gate-level authentication and has more communication overheads. Rubika method of authentication [6] is also designed to support client authentication only. Wei et al [7] ensure the authenticity of messages sent from one MR-job component to another. There is need of mutual authentication between MR components and MR infrastructure components. J. Zhao et al [8] supports the authentication of a client to MR application and authentication between a pair of domain specific MR components. Authentication layered model in [19] is implemented for providing mutual authentication between MR components.

B. Related Systems

Bao et al [14] proposed a joint emotion-topic model which is built by augmenting Latent Dirichlet Allocation with an intermediate layer for emotion modeling. The proposed model does not treat each term in the document individually and allows associating the terms and emotions via topics which are more flexible and has better modeling capability.

Tai et al [16] built Feeling Distinguisher system based on Supervised Latent Dirichlet Allocation, Latent Dirichlet Allocation and SentiWordNet methodologies for detecting a person's intention and intensity of feelings through the analysis of his/her online posts. The performance of FeD is about 1.08 to 1.18 folds that of SVM and sLDA.

Nie et al [17] implemented a text data crawler for mainstream online news websites, the modules of document preprocessing, document representation, and also integrated successful emotion analysis methods and provided the corresponding performance evaluation. SEAS automatically analyzed the emotions towards certain news articles and output the predicted emotions and probabilities of being classified into these emotion categories.

Lei et al [18] presented a lexicon-based approach towards social emotion detection. In their approach, the generated social emotion lexicon serves as the foundation of emotion detection, sentiment analysis, and opinion mining, as it can be used to discover what aspects the society likes and dislikes, to know the targets of each emotional state. Statistical results show a significant improvement in terms of accuracy for systems with the document selection process. Quantitative analysis on performance, the consideration of POS information does improve the accuracy on average, even though the corresponding improvement is statistically not too significant. Proposed method of generating the lexicon has a solid statistical foundation, and it outperforms the baselines significantly. We surveyed Sentiment Analysis, Emotion Mining, Text classification approaches in [19].

III. PROPOSED METHODOLOGY

Candidate registers for a specific organization using the web application. Candidate provides screen name of Twitter. Product Administrator access web application to fetch candidate's tweets from Twitter using a screen name. Alluxio is used as the main storage system which stores in-memory data. Apache's HDFS [10] is used as an underFS storage system. Candidate's CSV files, trained models, testing and training dataset is stored in Alluxio. Document classifier is built in Spark. Before training, documents are preprocessed and transformed using hashing trick. Features are extracted

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

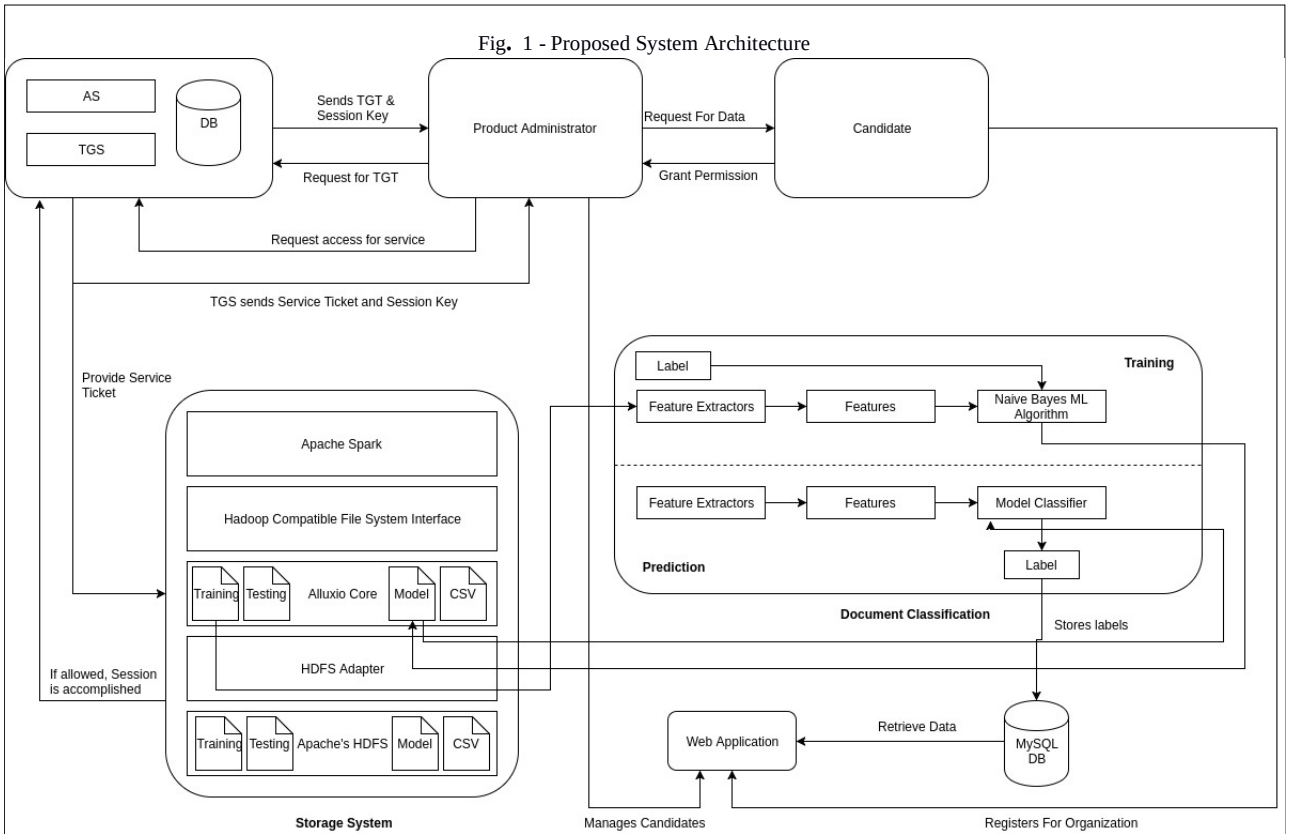
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

from training documents and feed it to the classifier for training. Spark's Mlib [9] is used for training Multinomial Naive Bayes, Multinomial Logistic Regression and Random Forest model which is stored in Alluxio.

Fig. 1 - Proposed System Architecture



For document classification, it access candidate's CSV file containing tweets and loads trained model with higher accuracy from Alluxio. It classifies each candidate's tweet into proper emotion and polarity classes. Predicted labels on each tweet is stored in MySQL database with the year, month and day. Profiling of candidates is done based on emotional and polarity score achieved. Candidates are compared based on emotional and polarity categories and their respective achieved score in the category. Kerberos is used for mutual authentication between storage system components and document classifier. It consists of Authentication Server, Database, and Ticket Granting Service. Product Administrator requests Key Distribution Center for a service ticket before submitting Spark job of document classification. If no valid ticket found, the operation is not permitted. With an only valid ticket, candidate's tweets are analyzed. After a certain period of time, the ticket is expired and Product Administrator has to request service ticket again.

Year-wise emotional and polarity scores for every category is formed to be time series dataset. Autoregressive Integrated Moving Average model is applied to time series dataset to predict next emotional and polarity scores for each category. Model is configured to operate with 1^{st} order auto-regressive, 0^{th} order moving average and no differencing to make time series stationary. Predictive Analytics powered by ARIMA model gives candidate's emotional and polarity scores of next five years.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

IV. EXPERIMENT AND RESULTS

The experiment is performed on single spark node with 8 cores and 8 GB RAM and four node spark cluster with 32 cores and 32 GB RAM. Overall execution time for training the model is recorded for a single node and multi-node formation.

A. Dataset

Dataset comprises of tweets from Twitter. It has to be collected for every candidate that needs to be assessed for behavioral assessment. There is significant latency and load on the server in fetching such information of candidate using Twitter API. After the dataset is collected, it needs to be stored in Alluxio. Movement of the huge dataset to storage layer requires additional I/O writes and communication overhead. But, by using the proposed system, writes operation for replication is significantly lesser [12]. For document classification, a training and testing dataset is required. Table 1 and 2 shows the number of training records containing documents related to specific emotional and polarity category respectively.

Emotion Category	**Training Records**
Anger	1572
Disgust	761
Fear	2839
Joy	8276
Love	216
Sadness	3853
Surprise	3912
Total	**21429**

Table 1: Emotion Dataset

Polarity	**Training Records**
Positive	12501
Offensive	12501
Negative	2007
Total	**27009**

Table 2: Polarity Dataset

B. Results and Discussion

This section provides the performance results of machine learning algorithms on emotional and polarity dataset. Multinomial Naive Bayes is configured to operate with smoothing constant set to 1. Random Forest is configured to use 3 trees, 32 bins, maximum depth of 4 and algorithm automatically chooses features subset strategy. Multinomial Logistic Regression is built for multiple classification with default settings. The emotional dataset is split into 10% testing and 90% training dataset. Multinomial Naive Bayes achieves higher accuracy, but Multinomial Logistic Regression achieves less weighted false positive rate than others.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Machine Learning Algorithms	Emotion Dataset				
	Accuracy	Weighted Precision	Weighted Recall	Weighted F1 Score	Weighted False Positive Rate
Multinomial Naive Bayes	53.22%	52.10%	53.22%	51.54%	17.23%
Multinomial Logistic Regression	45.39%	45.23%	45.39%	45.24%	16.02%
Random Forest	37.25%	19.87%	37.25%	20.30%	37.12%

Table 3: Performance Evaluation on Emotion Dataset

Polarity dataset is split into 30% testing and 70% training dataset. Machine learning algorithms are configured same as for emotion dataset. Logistic Regression achieves higher accuracy and less weighted false positive rate than others.

Machine Learning Algorithms	Polarity Dataset				
	Accuracy	Weighted Precision	Weighted Recall	Weighted F1 Score	Weighted False Positive Rate
Multinomial Naive Bayes	80.02%	80.04%	80.02%	80.02%	20.06%
Multinomial Logistic Regression	81.377%	81.371%	81.377%	81.372%	18.63%
Random Forest	57.31%	59.53%	59.31%	54.67%	42.69%

Table 4: Performance Evaluation on Polarity Dataset

Figure 2 and 3 gives performance evaluation of overall execution time with the different number of cores of machine learning algorithms on emotion and polarity dataset. As shown in Figure 3, there is a significant decrease in overall execution time of machine learning algorithms in 4-Node cluster formation with 32 cores.

Fig. 2 - Emotion Dataset

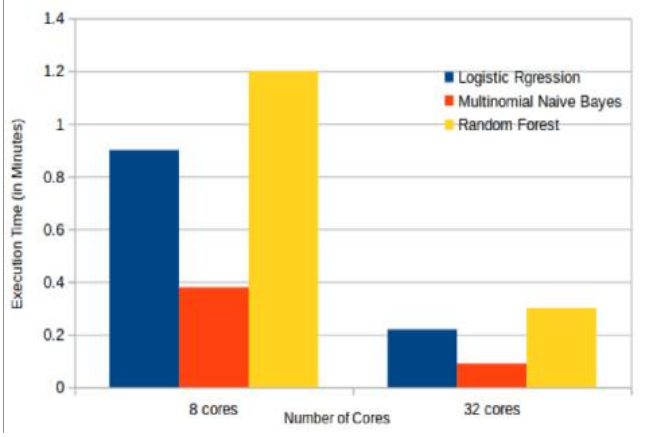
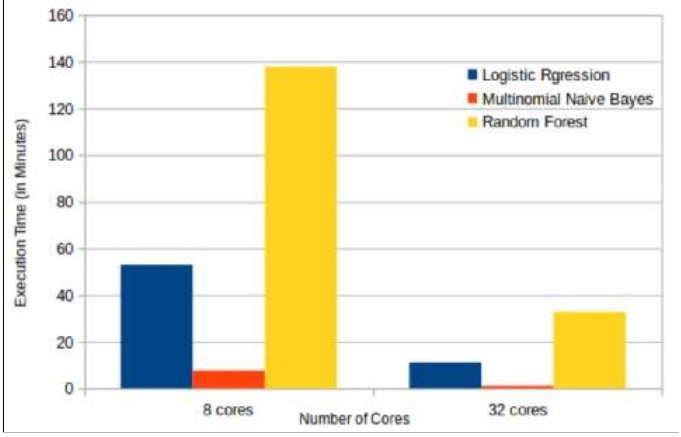


Fig. 3 - Polarity Dataset



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

V. CONCLUSION AND FUTURE WORK

The proposed system analyzes candidate's tweets for meaningful insights, assessing their behavior for their proper placement in organization's hierarchy. There are communication overhead and latency involved in retrieving candidate's tweets from Twitter and placing them in the storage system. The storage system uses Alluxio core services integrated with Hadoop framework, that allows minimal delay while writing data to the system by using the concept of lineage. The system also uses HDFS as an underFS storage system for persistence. There is minimal latency involved while moving data blocks from underFS storage layer to Memory Layer. Kerberizing Alluxio cluster provides a proper authentication between the storage system and Spark components. Without a valid ticket, the client is not authenticated to use any of the system's services. Proposed system improves the overall performance of the emotion and polarity classification with higher accuracy and less false positive rate.

The research can be extended to explores the relationship between behaviors and psychological theories to determine candidate's language style or social tendencies. The system only fetches 3200 tweets of a candidate for analysis. For more precise profile deviation, more tweets should be fetched for Behavioral Analytics.

REFERENCES

1. C.O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-Based Emotion Prediction", *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 579-586, 2005.
2. C. Strapparava and R. Mihalcea, "Learning to Identify Emotions in Text", *Proceedings of the 23rd Annual ACM Symposium on Applied Computing (SAC 08)*, pp. 1556- 1560, 2008.
3. R. Tokuhisa, K. Inui, and Y. Matsumoto, "Emotion Classification Using Massive Examples Extracted From The Web", *Proceedings of the 22nd International Conference Computational Linguistics (Coling 08)*, pp. 881-888, 2008.
4. Changhua Yang, Kevin Hsin-Yih Lin, Hsin-Hsi Chen, "Emotion Classification Using Web Blog Corpora", *Web Intelligence, IEEE/WIC/ACM International Conference on*, Fremont, CA, pp. 275-278, 2007
5. N. Somu, A. Gangaa, and V. S. S. Sriram, "Authentication service in Hadoop using one time pad", *Indian Journal of Science and Technology*., Vol. 7, pp. 5662, Apr. 2014.
6. S. Rubika, G. S. Sadasivam, and K. A. Kumari, "A novel authentication service for Hadoop in cloud environment", *Proceedings of theIEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, Oct. 2012, pp. 16.
7. W. Wei, J. Du, T. Yu, and X. Gu, "SecureMR: A service integrity assurance framework for MapReduce", *Proceedings of the Annual Computer Security Applications Conference*, Dec. 2009, pp. 7382.
8. J. Zhao, J. Tao, and A. Streit, "Enabling collaborative MapReduce on the cloud with a single-sign-on mechanism", *Computing*, Vol. 98, No. 1, pp. 5572, Jan. 2014.
9. Apache Spark website. [Online]. Available: http://www.spark.apache.org/, Accessed on: Feb, 2017
10. Apache Hadoop website. [Online]. Available: http://hadoop.apache.org/, Accessed on: Feb, 2017
11. Alluxio website. [Online]. Available: http://www.alluxio.org/, Accessed on: Feb, 2017
12. Haoyuan Li , Ali Ghodsi , Matei Zaharia , Scott Shenker , Ion Stoica, "Tachyon: Reliable, Memory Speed Storage for Cluster Computing Frameworks", *Proceedings of the ACM Symposium on Cloud Computing*, pp. 1-15, November 03-05, 2014, Seattle, WA, USA.
13. Kerberos website. [Online]. Available: https://web.mit.edu/kerberos/, Accessed on: March, 2017
14. Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu, "Mining Social Emotions from Affective Text", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 24, No. 9, pp. 1659-1661, Sept. 2012
15. I. Lahmer, N. Zhang, "Toward a Virtual Domain Based Authentication on MR", *IEEE Access*, Vol. 4, pp. 1662-1667, April 2016.
16. Chih-Hua Tai, Zheng-Han Tan, and Yue-Shan Chang, "Systematical Approach for Detecting the Intention and Intensity of Feelings on Social Network", *IEEE Journal of Biomedical and Health Informatics*, pp. 1-8, 2016.
17. Peng Nie, Xue Zhao, Li Yu, Chao Wang, Ying Zhang, "Social Emotion Analysis System for Online News", *12th Web Information System and Application Conference*, pp. 1-6, 2015
18. Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, Liu Wenyin, "Towards building a social emotion detection system for online news", *Future Generation Computer Systems*, pp. 439-447, 2014.
19. Sagar S. Patil, Pravin S. Game, "Sentiment Analysis, Emotion Mining and Authentication Methods in Hadoop: A Survey of Approaches", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 6, Issue 3, pp. 309-311, March, 2017.