IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Sentiment Analysis, Emotion Mining & Authentication Methods in Hadoop: A Survey of Approaches

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

Student, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India [1]

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

Abstract: With explosion of data, there is a need of finding new techniques to find patterns deep hidden for making key decisions. These patterns reveal significant information beneficial to organizations. Storing and efficient processing such “Big Data” for multiple organisations is a challenge. In this paper, a few approaches for analyzing big data for sentiment analysis and emotion mining are dicussed along with a few authentication models for processing frameworks to isolate analysis jobs of multiple users.

Keywords: Emotion Mining, Hadoop, Kerberos Authentication, MapReduce, Sentiment Analysis, Apache Spark, Apache Mahout.

I. INTRODUCTION

Sentiment Analysis refers to understanding behaviour of subject with respect to some topic. It is a computational study of people's opinions and emotions towards entities, events and their attributes. It determines overall polarity or emotional context to a document or event occurred. It uses text analytics to mine various sources of data. For analysis, dataset needs to be collected from most used social media sites like Facebook and Twitter. For emotion classification, some emotion categories are decided. After analyzing any text, it will be classified into one of the emotion categories. There is certain classification process that needs to be followed. It will be dicussed in Section II.

Data gathered from multiple sources comes in multiple formats. Data in multiple formats is processed to draw out certain conclusion. Unstructured data requires transformation to structured data for processing framework to analyze. There are big data technologies that are equipped with tools that can process and analyze data in various formats.

Emotion Mining is a process of extracting emotions from text. Studying emotions of people for an event occurred helps organisations make preemptive decisions. Emotions becomes ideal resource for servicing business and decision making. Word-emotion associations lexicons are widely used resources for analyzing emotions. In Section IV, a few emotion mining approaches will be dicussed.

Data is growing in nature, hence it is a challenge to store such data. Apache Hadoop's HDFS is one of the option for storing data for persistence. Data is valuable and storing all the data at one location is risky. HDFS provides replication of data blocks on distributed connected nodes achieving fault tolerance.

Analysis of big data takes longer time to draw out necessary conclusions. A faster approach is required that will speed up the computation by distributively perform local computation on set of connected commodity hardware and return the result to master node. This will significantly decrease overall execution time of big data application. A few technologies are dicussed in Section VII.

Authentication is the process of identifying a processing component based on a set of secret keys. Most of the processing frameworks do not have built-in authentication. Various components in framework can be compromised by range of attacks. This may leads to data leakage and malfunctioning of processing components. It is necessary to isolate jobs submitted by users that are handling sensitive data. A user needs to be authenticated first before submitting any job. A few authentication approaches are dicussed in Section VI.

Authorization is the process of giving individuals access to system processing objects based on their identity. A client should have a limited access to some of the Hadoop components. If user is given complete control of components, then altering configuration of components leads to erroneous results.

Apache Hadoop [15] is an open source, programming based framework that supports the processing and storage of extremely large datasets in a distributed computing. With Hadoop, it is possible to run applications on systems with number of commodity hardware nodes. Using its distributed file system, rapid data transfer rates is achieved. For analyzing huge dataset of textual data for classification, Hadoop is used for storage. Hadoop clusters are configured to use Kerberos Authentication Protocol [19] for secure communication between processing components.

DOI 10.17148/IJARCCE.2017.6369

II. PROCESS OF TEXT CLASSIFICATION

Text Classification is a process of assigning a document to one or more categories. Documents are classified according to their subjects and other attributes.

Any document to be classified needs to go through a certain classification process. But, as the data is big in size, it is necessary to modify the process to perform distributed operations compatible with processing framework, Hadoop.

Following are the stages involved in text classification:

A. Document Collection

A dataset containing number of documents needs to be collected. These documents will be classified to emotion categories. A dataset to train the classifier needs to be prepared. Training dataset will contain set of documents pre-classified to emotion categories.

B. Pre-Processing Documents

The first step of pre-processing documents is representing documents in <Text, Text> sequence format. Then sequence format is converted to <Text, VectorWritable> sequence file format containing term frequencies for each document. After preprocessing, set of unique terms in dictionary and TFIDF vectors is generated.

C. Indexing

Indexing is used to reduce the complexity of documents. It will be easier to handle, as documents are transformed from sequence file format to document vectors. TF-IDF is used as an weighting factor as it's value increase proportionally to the number of times a word appears in the document. It reflects how important a word is to a document in a dataset. Other methods in [6] are ontology representation for a document, sequence of symbols called N-Grams, multiword terms represented as vector components, Latent Semantic Indexing (LSI) and Locality Preserving Indexing (LPI).

D. Feature Selection

By representing documents in TF-IDF vector model, features, which are significant, are selected. It improves efficiency and accuracy of a text classifier. Subset of features which represents unique words from original documents are selected.

E. Classification

Documents can be classified by three ways: Unsupervised, Supervised and Semi-Supervised. It will be discussed more in Section V.

F. Performance Measures

Effectiveness of the classifier is evaluated. Therefore, performance of classifier is measured. Following are the measures for performance evaluation:

1) Accuracy: proximity of classified emotion to the true emotion.
2) Sensitivity: measures the proportions of emotion classes that are correctly identified.
3) Specificity: measures the proportion of negatives that are correctly identified.
4) Positive predictive value: are the proportions of positive results in statistics and diagnostic tests.
5) Negative predictive value: are the proportions of negative results in statistics and diagnostic tests.

III.SENTIMENT ANALYSIS APPROACHES

There are two main approaches in analyzing sentiments. Lexicon based approach determines collective polarity of a document by suming polarities of the individual words. Machine learning based approach by training a classifier with dataset containing documents pre-classified to certain polarity. Analysis is done on testing dataset which consist of documents which needs to be classified. With strong mathematical optimizations, a model is constructed from training dataset to make decisions. Sentiment Analysis is performed on three levels.

A. Document Level Analysis

It is performed on whole document and then determined what may be polarity of the document.

B. Entity Level Analysis

It finds precise sentiment on entities. It performs fine grained analysis.

C. Sentence Level Analysis

It finds what polarity a each sentence expressed. It is closely related to subjectivity classification.

In machine learning approach, there are two types of learning techniques. Supervised and Unsupervised Machine Learning Techniques. Different classification algorithms are used to correctly extract sentiment out of documents. A few of classification algorithms are discussed in Section V.

Sisi Liu et al. [24] combines K-Means Clustering and Support Vector Machines to analyze sentiments in emails. They also tried combining different machine learning algorithms but combining K-Means with SVM gives better accuracy.

M. Venugopalan et al. [25] applied hybrid model for sentiment analysis. The model was applied on Twitter dataset. Tweet specific features was extracted. Model used domain independent and domain specific lexicons. Average improvements was achieved in different domains.

Gang Li et al. [11] introduced clustering based approach for sentiment analysis. TF-IDF weights were calculated. Clustering result was improved by importing term scores and voting mechanism. It is more efficient than supervised learning.

DOI 10.17148/IJARCCE.2017.6369

IV. EMOTION MINING APPROACHES

Emotion Mining can be divided into three categories. The first category aims to extract valence of the text. It indicates polarity of emotions associated to it. Second category aims to determine whether text is subjective or factual. It determines if the text contains emotions or not. The third category aims to recognise intensity of emotions in the text.

A. Keyword Spotting

It is based on a lexicon or a dictionary grouping words that have emotional associations. It predict the emotions of the writer by identifying affective words from the text. Words are unambiguous and reflect clearly a particular emotion. It is simple to form word-emotion association and predict emotions from text using those associations. If sentence is very complicated then emotions are predicted poorly as this approach is based on individual words. It fails to uncover underlying emotions and intensity from the text. It is incapable of recognizing sentences without keywords. Syntax structures and semantics carry influences on emotions expressed. Therefore, linguistic information is needed to detect emotion more accurately.

B. Lexical Affinity

A probabilistic affinity is assigned to each word for a certain emotion. This approach is a bit more refined than keyword spotting. Like keyword spotting, this approach also performs poorly when sentence is over complicated.

C. Natural Language Processing

In this approach, machine learning algorithms are used to learn lexical affinities and word co-occurence frequencies. A large text corpus is used as training the classifier. Accuracy of predicted emotion depends upon trained classifier and corpus used for training. A poorly trained classifier will surely leads to wrong prediction of emotions. Considering social media network's domain, it is hard to set statistical rules as language used lacks proper structure.

D. Handcrafted Models

To mine emotions from text, a model uses deep understanding of the particular text. They are complex systems and their findings are difficult to generalize to other texts. A particular event or situation is examined. The feelings invoked by this particular event helps to determine emotional intensity in the text. This approach provides accurate prediction of emotion from over complicated sentence. It depends upon particular dataset used which comprises of real world knowledge.

Bao et al. [21] proposed a system to mine social emotions from affective text. The system finds out the relation between user generated emotions and online documents. It generates a set of hidden topics from emotions. Then it generates affective terms from each topic. A new joint emotion-topic model by augmenting Latent Dirichlet Allocation for modeling emotions was proposed. The generated model is effective and efficient in extracting the meaningful topics. It significantly improves the performance of social emotion prediction.

Yang Shen [22] have proposed a system which mines emotions from a blog. The dictionaries with weighted words, negative words, degree words and interjection words was defined. 23.8% blogs first sentence expresses the main idea while in 51.3% blogs last sentence expresses the main idea. System also calculated emotional index. It achieved accuracy rate of 80.6%.

Kamath S.S. et al. [23] discussed various approaches in the field of emotion mining and sentiment mining. They proposed a model for analyzing bias in online content. The model calculates credibility scores for articles based on sentiment difference between subtopics and between websites.

V. TEXT CLASSIFICATION APPROACHES

Machine learning is a method of analyzing data to build a automatic analytical model. We use Machine Learning Algorithms to iteratively learn from data and find hidden insights.

Mainly, there are two types of Machine Learning Algorithms:

A. Unsupervised Machine Learning

When dependent variable is unknown, this type of machine learning is preferred. It is mainly used for clustering in different groups.

Following are a few approaches of Unsupervised Machine Learning –

1) K. Nearest Neighbor:

Objects are classified by votes [20]. Voting is done for several labeled training dataset with their smallest distance from each object. This method is simple and performs well even in handling the classification tasks with multi-categorized documents. But it takes more time for classifying objects when training dataset given is large.

2) Neural Network:

Input units represent terms and the output units represents the categories. The weights on the edges connecting units represent dependency. For classifying a given test document, its term weights are loaded into the input units; the activation of these units is propagated forward through the network, and the value of the output units determines the categorization decisions.

B. Supervised Machine Learning

This type of Machine Learning Algorithm focuses on dependent variable and to be predicted from set of independent variables. Using our training dataset, it classifies documents to one of the emotion and subject

DOI 10.17148/IJARCCE.2017.6369

categories. Training phase continues till our model reaches a higher level of accuracy.

Following are the approaches of Supervised Machine Learning –

1) Support Vector Machines:
A SVM model represents points in space, mapped points of the separate categories are divided by a clear gap that is as wide as possible. New points are then mapped into same space and predicted to belong to a category based on which side of the gap they fall. In addition to linear classification, SVM also efficiently performs non-linear classification by mapping their inputs into high dimensional feature spaces. For learning text classifier [11], we have to deal with many features. SVM have the potential to handle these large feature sets. But it only makes to have to a featues that are relevant. SVM considers only the features that are truly relevant. The idea of SVMs is to find such linear separators as most text categorization problems are linearly separable. SVM are well suited for problems with dense concepts and sparse instances.

2) Multinomial Naive Bayes:
A probabilistic classifier [12] which uses Bayes's theorem with naive independence assumptions between the features. It is highly scalable and requires number of parameters linear in the number of variables in a learning problem. Classifier is based on bag of words model. A document is formed by drawing words from a multinomial distribution. Documents are represented in integer vector, elements indicating frequency of occurrence of a word. In pre-processing, multiple occurrences of words are taken into account. It is works well with longer documents and large dataset.

3) Decision Trees
Internal node represents the label as term and leaf nodes represent corresponding class labels. Tree classifies the document by running through structure from the root to until it reaches a certain leaf node. It is [9] inefficient because of frequent swapping of training tuples. Swapping is required because most of the training data will not fit into memory decision tree construction.

4) Random Forest
Random forest is a popular classification method which randomly select subspace of features at each node to grow branches of a decision trees, then to use bagging method to generate training data subsets for building individual trees, finally to combine all individual trees to form random forests model. Baoxun Xu et al. [10] explains that during building of forest, topic-related or informative features would have the high chance to be missed, if randomly small subspace from high dimensional text data is selected. Therefore, weak trees will be created from these subspaces. Thus, it will have a large likelihood to make a wrong decision.

VI. AUTHENTICATION APPROACHES

A. Authentication Methods
Multiple clients submit their respective jobs for processing. Before submitting any jobs, a client needs to get authenticated by Authentication Server. Somu et al. [14] method is symmetric key based. It uses single authentication factor, supports only gate-level authentication and has more communication overheads. Rubika et al. [6] is also designed to support client authentication only. Wei et al. [5] ensures 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. [4] supports the authentication of a client to MR application and authentication between pair of domain specific MR components.

B. MapReduce Layered Authentication Model
Hadoop does not have any in-built security mechanism. Therefore for security purposes, a mutual authentication among hadoop components and server nodes is developed. I. Lahmer et al. [3] included three layers of security. First one is mutual authentication among server nodes. Second one is authenticating hadoop components to server node. Both of them can be accomplished by implementing Kerberos Authentication Protocol. Third one is mutual authentication among hadoop components. It is required when there are multiple mapreduce jobs in execution by multiple product administrator and isolation to job operation is provided. VDAF, a layered authentication model provides authentication for multiple cross domain mapreduce jobs. It is achieved by assigning two types of identifiers to hadoop components. Static identifiers are carried by MR components that are shared by more than one job. Dynamic identifiers are assigned when jobs are created and they are destroyed when execution is finished.

VII. TECHNOLOGIES

Big data processing requires certain technologies to store, analyze and visualize results. Serial processing of every record takes longer time to produce meaningful conclusions. A distributed processing approach on commodity hardware should be considered.

Following technologies supports Machine Learning Algorithms and used for writing Big Data Application –

A. Apache Mahout
It is used to build an evironment for creating scalable performant machine learning application. Sklearn, a python machine learning library that supports number of classification and clustering algorithms can also be used. But, as the classification sequential operations take a lot of time for execution, Apache Mahout [16] is used. Mahout provides useful tools like seqdirectory and seq2sparse for pre-processing documents. Both tools runs distributively as a mapreduce jobs on top of Hadoop. Mahout provides

DOI 10.17148/IJARCCE.2017.6369

implementation of machine learning algorithms for text classification and sentiment analysis.

B. Apache Spark

Spark [17] provides fast cluster computing system. It's high level APIs in Java, Python, Scala, Python and R are used. It also uses Hadoop client's libraries for performing operation on data stored in HDFS.

Spark system in cluster mode is deployed to take advantage of commodity hardware by using Hadoop YARN as cluster manager. Spark with Mahout APIs can be used for calculating term frequencies, generating TFIDF vectors, writing and reading from HDFS and Alluxio and classification. By using Spark, more records in less time is classified efficiently.

C. Apache Hadoop

Hadoop's HDFS [15] stores huge amount of training data in HDFS for distributed processing. Mahout API's can access data stored in HDFS for processing. It provides scalable environment and fault-tolerant file systems.

D. Alluxio

It provides memory centric design and unified namespace for different storage system. Alluxio [18] improves processing speed for big data applications while providing a common interface of data access. Haoyuan Li et al. [1] explains how Alluxio speeds up read and write throughtput.

E. Kerberos

To provide a level of security among Hadoop components like Resource Manager etc. Kerberos [19] sets up a Key Distribution Center. KDC has three main components. Kerberos Database which stores users and services identity. Authentication server resides as a separate physical server.

It issues tickets upon initial request. Ticket Granting Service responsible for providing service tickets. By using these three components, communication is secured between core components of Alluxio, Spark and Hadoop.

VIII. CONCLUSION

User's social media presence and content published in various sites can be mine for opinions and emotions. In this paper, a few approaches of sentiment analysis and emotion mining are discussed. Machine learning approaches for sentiment analysis are effective and provides accurate predictions than lexicon based approaches. To mine emotions from text, a deep understanding at sentence level is preferred.

A model should understands semantics and syntax of sentences for accurate prediction of emotion labels. Layered approach establishes authentication levels at server nodes and Hadoop components.

ACKNOWLEDGMENT

I sincerely convey my gratitude to **Mr. Vijay Kumbhar**, Co-Founder and VP Engineering at Webonise Labs Pvt. Ltd. for his constant support, providing all the help, motivation and encouragement from beginning till now to make this work a perfect one.

REFERENCES

[1] 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.

[2] 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

[3] I. Lahmer, N. Zhang, "Toward a Virtual Domain Based Authentication on MR," In IEEE Access, Vol. 4, pp. 1662-1667, April 2016.

[4] 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.

[5] W. Wei, J. Du, T. Yu, and X. Gu, "SecureMR: A service integrity assurance framework for MapReduce," in Proc. ACSAC, Dec. 2009, pp. 7382.

[6] S. Rubika, G. S. Sadasivam, and K. A. Kumari, "A novel authentication service for Hadoop in cloud environment," in Proc. IEEE Int. Conf. Cloud Comput. Emerg. Markets (CCEM), Oct. 2012, pp. 16.

[7] Peng Nie, Xue Zhao, Li Yu, Chao Wang, Ying Zhang, "Social Emotion Analysis System for Online News", 2015 12th Web Information System and Application Conference.

[8] 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.

[9] Vandana Korde, C Namrata Mahender, "Text Classification And Classifiers: A Survey", International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012, pp. 87-89

[10] Baoxun Xu, Xiufeng Guo, Yunming Ye, Jiefeng Cheng, "An Improved Random Forest Classifier for Text Categorization", Journal Of Computers, Vol. 7, No. 12, December 2012, pp. 2913

[11] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features", pp. 3-4

[12] Hiroshi Shimodaira, "Text Classification using Naive Bayes", Learning and Data Note 7, Informatics 2B, 10 February 2015, pp. 8-9

[13] Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, Liu Wenyin, "Towards building a social emotion detection system for online news", Future Generation Computer Systems, 21 October 2013, pp. 438-448.

[14] N. Somu, A. Gangaa, and V. S. S. Sriram, ''Authentication service in Hadoop using one time pad,'' Indian J. Sci. Technol., vol. 7, pp. 56–62, Apr. 2014.

[15] Apache Hadoop website. [Online]. Available: http://hadoop.apache.org/

[16] Apache Mahout website. [Online]. Available: http://mahout.apache.org/

[17] Apache Spark website. [Online]. Available: http://www.spark.apache.org/

[18] Alluxio website. [Online]. Available: http://www.alluxio.org/

[19] Kerberos website. [Online]. Available: https://web.mit.edu/kerberos/

[20] Pratiksha Y. Pawar and S. H. Gawande, "A Comparative Study on Different Types of Approaches to Text Categorization", International Journal of Machine Learning and Computing, Vol. 2, No. 4, August 2012, pp. 423.

DOI 10.17148/IJARCCE.2017.6369

[21] Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han and Yong Yu, "Mining Social Emotions from Affective Text," in Knowledge and Data Engineering, IEEE Transactions on, Vol. 24, No. 9, pp. 1658-1670, Sept. 2012.

[22] Yang Shen, Shuchen Li, Ling Zheng, Xiaodong Ren, Xiaolong Cheng, "Emotion mining research on micro-blog," in Web Society, 2009. SWS '09. 1st IEEE Symposium on, pp. 71-75, 23-24 Aug. 2009.

[23] S. Kamath, A. Bagalkotkar, A. Kandelwal, S. Pandey, K. Poornima, "Sentiment Analysis Based Approaches for Understanding User Context in Web Content," in Communication Systems and Network Technologies (CSNT), 2013 International Conference on, pp. 607-611, 6-8 April 2013.

[24] Sisi Liu, Ickjai Lee, "A Hybrid Sentiment Analysis Framework for Large Email Data," in Intelligent Systems and Knowledge Engineering (ISKE), 2015 10th International Conference on, pp. 324-330, 24-27 Nov. 2015.

[25] M. Venugopalan, D. Gupta, "Exploring sentiment analysis on twitter data," in Contemporary Computing (IC3), 2015 Eighth International Conference on, pp. 241-247, 20-22 Aug. 2015.

[26] Gang Li, Fei Liu, "A clustering-based approach on sentiment analysis," in Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on, pp. 331-337, 15-16 Nov. 2010.

DOI 10.17148/IJARCCE.2017.6369