RESEARCH REVIEW JOURNALS
Volume-04
Issue-04
April -2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com [UGC Listed Journal]

Integration of Sentiment Analysis engine with Big Data with Lexical Based approach

[1]Kanwaldip Kaur & [2]Dr. Rajan Manro

[1]Research Scholars (Deptt. of Computer Science)
[2]Deptt. of Computer Science, Khanna(Punjab)

ARTICLE DETAILS

Article History
Published Online: 15 April 2019

Keywords
Big Data, MongoDB, Ontology, Lexical.

ABSTRACT

In today's environment there are various applications which are producing large amount of data. These data items requires analysis for use this data for certain useful purpose in the industries. These processed data items can useful for understanding the behavior of the person present on the social media places. In current research paper data belongs to various social media sites is being processed using Hadoop and processed data will be stored into the MongoDB. Later on using certain simulator data will be processed to generate the comparative analysis. The proposed approach based on the Lexical based analysis is applied for the data analysis and classification purpose. The data extracted after lexical analysis will be compared to the ontology of the generated positive and negative words. The proportional positive and negative classification is done for the analyzed data. This will help in identifying the sentiments of the people about certain items for which the data is being shared. The proposed approach is compared to the existing approach based on different parameters like Accuracy, Time and the Error Rate. The proposed approach is giving better results. That means provides better accuracy, less time for the computation and less error rate.

1. Introduction

Big Data mining may well be a way to go looking out important patterns from the on the market text documents. Text mining, jointly remarked as text processing, is that the strategy of account high-quality information from text. 'High quality' in text mining usually refers to some combination of connectedness, novelty, and power [1]. High-quality information is usually derived through the assembly of patterns and trends extracted or evaluated through the means like applied mathematics pattern learning. Text mining usually includes the strategy of structuring the input text (usually parsing, along with the addition of some derived linguistic choices and additionally the removal of others, and ulterior insertion into a database), account patterns within the structured information, and ultimately analysis and interpretation of the output [3].

1.1 Stages of Text Mining Method

Text mining methods have been utilized in the versatile applications, ranging from the data retrieval to the natural language processing applications. The Text mining application requires the multiple steps to be executed in the particular arrangement, which is shown in the following steps:

1. **Data Retrieval** systems establish the documents in a very assortment that match a user's question. The foremost acknowledge IR systems are search engines like google that establish those documents on the globe wide net that are relevant to a collection of given words.
2. **Natural Language Process** (NLP) is one amongst the oldest and most troublesome issues within the field of computing. It's the analysis of human language in order that computers will perceive natural languages as humans do. This is usually done using the annotation documents with data like sentence boundaries, part-of-speech tags, parsing results, which might then be browse by the data extraction tools.
3. **Data Methoding** (DM) is that the process of characteristic patterns in massive sets of knowledge. The aim is to uncover antecedently unknown, helpful information. Once employed in text mining, DM is applied to the facts generated by the data extraction section and places the results of our DM method into another information which will be queried by the end-user via an acceptable graphical interface. The info generated by such queries may be delineated visually.
4. **Data Extraction** is that the method of mechanically getting structured knowledge from an unstructured language document.

1.2 Major Tasks of Text Data and Social Sites Data Classification

In general, new classification tasks may be classified into 2 categories: descriptive data processing and prophetical data processing. The previous describes the info set in laconic outline manner and presents attention-grabbing general properties of information. An information mining system might accomplish one or a lot of the subsequent data processing tasks.

- Category Description: Class description provides a laconic and account of assortment of information and distinguish it from others. The account of assortment of information is named category characterization, the comparison between 2 or a lot of collections of information is named comparison or discrimination.
- Association: Association is discovery of association relationships or correlations among a group of things. There are varied association analysis algorithms like

Apriori search, mining multiple level, multi dimensional association, mining association for numerical.

- Classification: Classification analyzes a group of coaching information (a set of object whose category label is known) and constructs a model for every category supported the options within the information. A choice tree or set of classification rule is generated by such a classification method. There are several classification technique developed within the field of machine learning, static, database, neural network.
- Prediction: This mining performs predicts the attainable price of some missing information and also the price distribution of certain attributes during a set of objects. It involve the finding of set of attributes relevant of the attribute of interest and predicting the worth distribution supported set of information almost like choose object.
- Agglomeration: Clustering analysis is tool established clusters embedded in information wherever a cluster may be a assortment of information object that's almost like each other. Similarity may be such that by user of consultants.
- Time Series Analysis: statistic analysis is to research giant set of your time series information to seek out bound regularities and attention-grabbing characteristics, together with rummage around for similar sequences and sub sequences, mining consecutive patterns, regularity, trends and deviation.

1.3.Machine Learning based Text Classification

1.3.1 Machine learning primarily based text classification contains quantitative approaches to car maize learning method that uses machine learning algorithm. most well supervised learning techniques for text classification area unit.

1.3.1.1. Rocchio Algorithmic rule(RAR)

Different words with similar meanings during tongue are termed as semantic relation. Semantic relation may be self addressed by purification the question or document victimization the connection feedback technique. The user provides feedback that indicates relevant material relating to the particular domain space.

The user ask the straightforward question and also the system generates initial leads to response the question. the user marks the retrieved results as either relevant or tangential. Numerous orthography correction techniques may be used at the once of computation and time interval, like hashing based mostly and context sensitive orthography correction techniques illustrated by author.

1.3.2.1 Decision Trees and Support Vector Machine

Relationships, attributes and categories in metaphysics may be structured hierarchically as taxonomies. The process of constructing lexical metaphysics by analyzing unstructured text is termed as metaphysics refinement. Totally different algorithms of call tree are used for classification in several application areas, Like Monetary analysis, astronomy, biological science, and text mining. Low range of relevant options during a class tree could cause poor performance in text classification. This idea has been explained by author.

It is employed to analysis of knowledge in classification analysis. In distinction to alternative classification strategies, SVM algorithmic rule uses each positive and negative knowledge sets to construct a hyper plane that separates the positive and negative data.

1.3.2.2 Artificial Neural Networks

Artificial Neural Networks is the distributed process system specifically galvanized by biological neural systems. The network contains of an outsized range of extremely interconnected process parts operating along to unravel any specific drawback. Owing to their tremendous ability to extract data which has meaning from an enormous set of information, neurons are designed fro specific application areas, like pattern recognition, features extraction, and noise reduction.

There are two basic classes of learning strategies employed in neural networks:(a) supervised learning and (b) unsupervised learning. In supervised learning, the ANN get trained with assistance of a group of inputs and needed output pattern provided by external skilled or an intelligent system. There are different kinds of supervised learning and that include:(a) Back propagation and(b) changed back propagation neural networks. In unsupervised learning , the neural network tends to perform cluster by adjusting the weights supported similar inputs.

2. Literature survey

SHAHID SHAYAA(2018) et. al: this paper focused on sentiment analysis for the big data. opinion mining and sentiment analysis for the Big data categories the people sentiment into different classes. Each class represents the people mood for the current aspect under consideration. This paper is focused on technical and non technical aspect of the system for the opinion mining.

Swati Redhu(2018) et. al: This paper focuses on the process of opinion mining and sentiment analysis for the Big data. The data collected is from different sources like social media, e-commerce, blogs, social media. This paper focuses on data acquisition, data pre-processing, Normalization and feature extraction and finally representation. This paper focuses on the different methods and techniques for the sentiment analysis and opinion mining.

DAVID ZIMBRA (2018) et al: This paper focuses on the process of classification for the sentiment analysis for different classes of the twitter data into different classes. The success rate of different classification is around 70%. 28 different commercial and academic system classifies different datasets into different classes. Paper focuses on the different trends and techniques for improving the results.

M.Trupthi(2017)et. al: This paper is focused on sentiment analysis of large number of reviews or tweet data on various social media sites. Large amount of data is being processed by

the processing engine. This processing engine can be Hadoop etc. This can process the large amount of data to extract the useful patterns. Sentiment analysis in current paper is based on features.

K. AMAROUCHE et al. (2016)numerous corporations in today's atmosphere square measure supported numerous natural philosophy market places. These market places as like net, and TV tec. They shares their contents through social networking sites. folks sell their merchandise on the web. therefore folks opinion concerning their product matters. as a result of negative reviewcan enforce merchandiser to seem at their merchandise therefore the product improvement. Even to took back their product from the market. They collects the review knowledge, analysis it and so soon method this data to extract helpful analysis.

C. Bucur et al. (2016) in line with this paper merchandiser or analyzer is to analysis the opinion expressed by the folks on touristy. The collected knowledge is to be summarized and classified in pre outlined dataset classification. With in the paper the potency of algorithms square measure analyzed victimization text mining domain specific measures.

A. Severyn et al. (2016) There needs sturdy model mistreatment that great deal of reviews knowledge by users will be sub setted and summarized for higher form of analysis. This sturdy model is specified it will handle great deal of knowledge, this knowledge even will have noise. This analysis is predicated on opinion mining on the information created in You Tube connected videos. This analysis is predicated on proposing a strong shallow grammar structure (STRUCT) that adapts well once tested across domains.

3. Proposed algorithm

Step1 Collect the Social Media data specially in relation to Tweeter data.
Step2 Collect the ontology of various positive and negative words.
Step3 Convert the Tweeter data into individual words using lexical based approach.
Step4 Identifying the number of positive and negative words into the Tweeter data.
Step5 Categories the reviews into two sub categories like positive and Negative.
Step6 Identify the collective analysis for all the two sub categories.

4. Existing algorithm

Step1 Acquisition of data related to Tweeter.
Step2 Identifying the three categories like Positivity, Negativity and Objectivity.
Step3 Scores are allocated by passing the words through sentiwordnet.
Step4 Specify the categories of value range. If total scores is less than 5 then it is considered negative review, if is greater than 5 then it will be considered as positive.
Step5 In last phase analysis is performed over to the scores values.

5. Ranking index algorithm

1. Acquire the Twitter data from the net supply or native supply
2. Extract the metaphysics technique primarily based keywords from the given tweet text
3.Apply the keyword matching and weight calculation mistreatment the supervised technique with the precise class primarily based list matching technique
4. Construct the keyword matching matrix mistreatment the pre-defined weight lists keep within the SRD (Sparse Ranking Data).
5. Restate the steps three and four iteratively for all tweet texts

In the Ranking formula, all coaching samples were used for coaching, that is, whenever the tweet classification sample or take a look at sample has to be verified, analyzed and classified, it's necessary to calculate similarities between that sample and every one documents within the coaching sets, so select Ranking with word samples that have largest similarities. because of numbers of calculation taken between the take a look at sample and every one the coaching samples, the normal technique of Ranking has less machine quality. To beat the quality, this paper introduced combination Ranking formula with a cluster technique.

6. The ranking algorithm for populating ranking index on the basis of word list matching

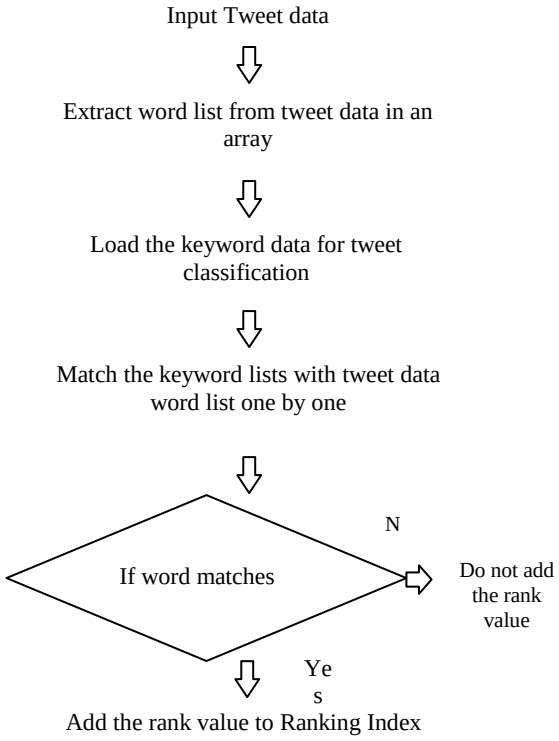

Fig. 1 Flowchart

7. Parameters Taken

1. Accuracy.

2. Time Based Analysis.
3. Error Based Analysis

8. Results

For proposed approach data is collected from different sources like Tweeter, Facebook and various other reviews sites involved in collecting the reviews data about the general goods the people purchase from the shoping web sites. The collected data will be processed by Hadoop and stored in MonDB. Later on using processing simulator data collected will be processed for the analysis purpose.

8.1 Error Rate

Proposed appraoch using lexical based analysis for the data collected from various social media platforms.

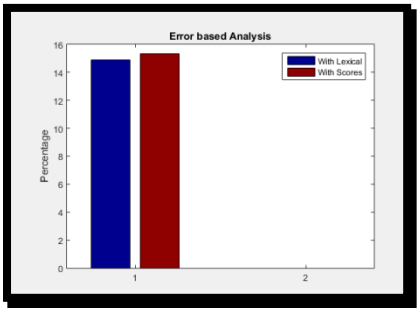

Fig. 2 Error Rate Comparison

The proposed approach shows less error rate compared to the score based technique. the proposed approach with lexical based sepration and comparing with the builded Ontology produces better results compared to the existing scores based technique.

8.2 Accuracy Comaprison

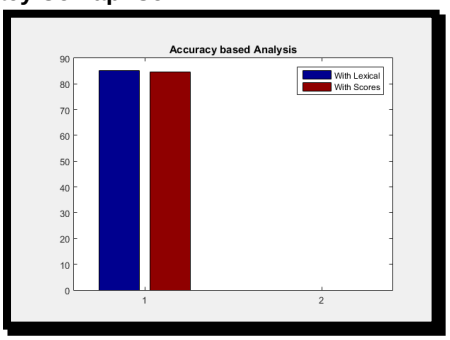

Fig. 3 Accuarcy Comparison

The proposed appraoch has better results in terms of accuarcy compared to the existing scores based appraoch.The proposed technique has accurate positive and negative sentoment extraction.

8.3 Time Analysis

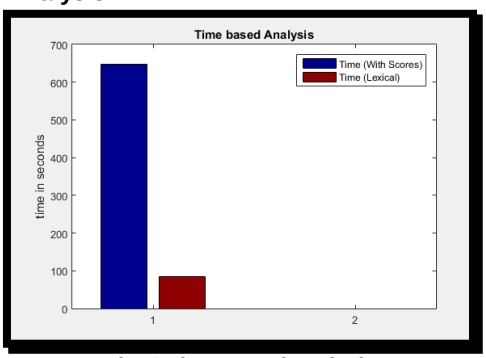

Fig. 3 Time Based analysis

The proposed technique based on Lexical based analysis is providing solutions with less time compared to the existing scores based technique. that means computationally proposed approach is more time efficient compared to the existing technique.

9. Conclusion

Various applications now a day's people are using are producing large amount of data. Processing this data will be required to drive the analysis for the decision making purpose. The sentiment analysis for the data of social media sites can be useful for the knowing the behavior of the person. This way the existing scores based technique mark the score value of the words extracted from the reviews data. If score value is greater than 5 then the sentiment will be considered positive and if scores value is less than 5 then the sentiment value will be considered negative. For enhancing the performance of the analysis lexical based approach is used for the reviews data for knowing the sentiment of the persons. In this approach first the reviews data based on separator will be sub divided into various individual words. The words are later on compared to the ontology of the positive and negative words. This proposed approach provides highly efficient results. The efficiency is compared on the basis of various parameters like error rate, accuracy, and time of evaluation. In all the aspect the proposed lexical based approach is providing better results in all the parameters.

10. Future Work

Currently the lexical based approach is applied on the data collected from various social media sites. Sentiment analysis for the social media data can helps in identifying the mood of the people regarding specific type of item. In future this work can be enhanced by using more specified ontology and also better approach for better efficiency.

References

1. Shahid Shayaa , Noor Ismawati ,Jaafar, Shamshul Bahri, Ainin Sulaiman, Phoong Seuk Wai, Yeong Wai Chung , Arsalan Zahid Piprani, And Mohammed Ali Al-Garadi, "Sentiment Analysis Of Big Data: Methods, Applications, And Open Challenges",Ieee,Vol. 6 2018.
2. Swati Redhu, Sangeet Srivastava1, Barkha Bansal , Gaurav Gupta," Sentiment Analysis Using Text Mining: A Review", International Journal On Data Science And Technology,Vol. 4(2),Pp: 49-53, 2018.
3. David Zimbra, Ahmed Abbasi, Daniel Zeng, Hsinchun Chen," The State-Of-The-Art In Twitter Sentiment Analysis: A Review And Benchmark Evaluation", Acm,Vol. 9, 2018.
4. M.Trupthi,Suresh Pabboju,G.Narasimha,"SENTIMENT ANALYSIS ON TWITTER USING STREAMING API",IEEE,2017.

5. Kamal AMAROUCHE, Houda BENBRAHIM, Ismail KASSOU," Product Opinion Mining for Competitive Intelligence",vol. 73,pp. 358 – 365,2015.
6. Cristian Bucur," Using Opinion Mining Techniques in Tourism",vol. 23,pp.1666-1673,2015.
7. Jumayel Islam, Zubair Azami Badhon and Pintu Chandra Shill," An Effective Approach of Intrinsic and Extrinsic Domain Relevance Technique for Feature Extraction In Opinion Mining",vol. 1, pp.1269-75,2016.
8. Shahab Saquib Sohail, Jamshed Siddiqui, Rashid Ali," Umw: Amodel For Enhancement In Wearable Technology Based On Opinion Mining Technique",vol. 1, pp.46-52, 2015.
9. Shoiab Ahmed, Ajit Danti," A Novel Approach for Sentimental Analysis and Opinion Mining based on sentiwordnet using Web Data",vo. 1 pp.15-20,2015.
10. Dhanalakshmi V., Dhivya Bino, Saravanan A. M.," Opinion mining from student feedback data using Supervised learning algorithms",vol. 1pp. 84-97,2016.
11. Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina , Barbara Plank , Katja Filippova ," Multi-lingual opinion mining on youtube",vol. 4,pp.45-54,2015.
12. Agarwal, Sonali, G. N. Pandey, and M. D. Tiwari. "Data mining in education: data classification and decision tree approach." International Journal of e-Education, e-Business, e-Management and e-Learning 2, no. 2 (2012): 140.
13. Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. "Sentiment analysis in the news." arXiv preprint arXiv:1309.6202 (2013).

