INTERNATIONAL EDUCATIONAL APPLIED RESEARCH JOURNAL (IEARJ)
Volume 03, Issue 07, July 2019
E-ISSN: 2456-6713

DATA PREPROCESSING IN SENTIMENT ANALYSIS USING TWITTER DATA

T. Nikil Prakash[1], Dr. A. Aloysius[2]

[1]*Research Scholar, Department of Computer Science, St. Joseph's College, Trichy, India.*

[2]*Assistant Professor, Department of Computer Science, St. Joseph's College, Trichy, India.*

Abstract: ***Data preprocessing is an important tool for Data Mining (DM) algorithm. Twitter data is an unstructured data set it is a collection of information from people entered his/her feelings, opinion, attitudes, products review, emotions, etc. This type of information is growing day by day in the internet. May companies want to analyze customers opinions which like the product and the services. The Proposed work to analyses the twitter trending information and collect various different information form the users. It improves the accuracy of Twitter data. This work easy to identify the people reaction or opinion. Additionally, improve the better performance for data preprocessing tool.***

Keywords: *Twitter, Data preprocessing, Sentiment analysis, Data cleaning, Data preparation.*

INTRODUCTION:

Big data preprocessing is a crucial task for many researchers, administrators, organizations and companies to collecting the data and analyzing the huge amount of specific data or information [1]. The different types of data or similar data are collected from different types of web sources and places. That data can be due to improperly measure in the term of noisy data, missing data, wrong data or inconsistent data. If the inequality data provide the wrong results and wrong conclusions in the data analysis, pattern reorganizations and decision making. The unique challenges are handled in mixed variety data and unstructured data order to be preprocessed into structured data or ordered data representations.

Data preparation is highly recommended of many reasons such as datum quality or database quality, process of data analysis, possibility of related algorithms to apply for removing noisy data and missing data and increase the data reliability that is high-quality data models require high quality of data [2].

Sentiment analysis involved identifying a given text of content by first preprocessing it to detecting stop words and symbols, etc. and then checking the subjectivity contents. The getting the opinion content polarity is determined either on machine learning methods and lexical based methods. Sentiment categorizes the content into positive or negative and or neutral. SA makes the use of knowledge in the term of context-dependent, for example, some single words gave multiple meaning in the given word. It can be solved by applying for proper context. If the proper context to increase the accuracy of sentiment classification in the knowledge of context to be used [3].

LITERATURE REVIEW:

Stamatios Aggelos. N et al [4] proposed a data preprocessing algorithm for predictive data mining algorithms. Its enable us to knowledge discovery form the large dataset. The DM algorithms are unreliable data to be received or noisy data cannot provide better results. The most well-known and widely used up-to-date algorithms provide the data preprocessing steps. The performance for all other frameworks to be compared. The DM algorithm handled the quality of data its too large or noisy values to be contained. Moreover, various classification algorithms are applied to similar problems.

Salvador Garcia et al [5] proposed data preprocessing methods, characteristics, and approaches. They communicate big data and data processing throughout all methods and technologies and include the state of art in big data. Additionally, it focuses on big data framework development, for example, Hadoop, spark, and flink. These applications and methods are learning new paradigms. They described big data preprocessing key issues and covered big data families of data processing that is feature selection, imperfect data, imbalanced learning, and instance reduction. Moreover, they developed in bid data preprocessing frameworks.

A. Asbrino et al [6] proposed semi-supervised SVM classification in data preprocessing. In SVM semi-supervised models contribute labeled and unlabeled data optimization, this model finding the good quality of separate hypo plane. These approaches are two types that are mirror integer linear programming problems and continuous optimizations problem. Both problems are solved to very hard and increased the computational model. It's reduced the number of unlabeled points and increasing performance classification.

Matthew J. Denny et al [7] proposed unsupervised learning methods for political science text data research and preprocessing decisions. They introduced statistical procedure and software that examines finding the sensitivity. This approach understanding the researchers'

problem by providing a characterization of variability changes in preprocessing. These techniques analyze the sensitivity of the researcher's results in preprocessing and take decisions. The comparative of data preprocessing in different specifications in the relative documents. This approaches based on theoretically and it is decreasing the risk for researchers.

Vivek Kumar et al[8] proposed source language sentence extraction (SLSE) framework model for a language translator. SLSE is creating a bilingual dictionary, N-grams, inverse term index, etc. SLSE is a training data set so it's a very difficult task for building model framework. The sentence selection has been described in well-defined function and the sentence has been extracted from the frequency of each generated query.

Shichao Zhang et al [9] proposed data preparation and data cleaning for data mining using machine learning systems. Data preparation is very important for research works and data preparation found the critical issues in the dataset. It described the possible directions of data preparations and delivered many challenges and issues for data preparations. The data preparation is very important criteria for data mining related algorithms because the real data is improving and provide the system's performance is a high quality of data. Moreover, the quality of the data produces concentrative patterns.

S. B Kotsiantis et al [10] proposed various algorithms for data preprocessing techniques to improve the best performance for the data set. It represents the quality of instance data. The irrelevant and redundant data provide noisy and unreliable data then the phrase is more difficult to form the knowledge discovery. The machine learning problems are also known as data preparation and processing time to be considered in the amount of data. The data preprocessing steps described machine learning algorithms. This algorithm produced less accurate and less understandable results or anything fail to discover. The preprocessing steps resolve several steps that include noisy data, redundancy data, and missing data.

Table 1: The comparison results of existing work.

Method	Types of approaches	Data source	Responsibility	
			Advantages	**Disadvantages**
S. B Kotsiantis et al	Machine learning	Twitter	Resolve noisy data, missing data, and redundancy	Less accurate Less understandable
Stamatios aggelos. N et al	Data mining	Twitter	Handled large data	Cannot provide better results
Salvador Garcia et al	Lexicon and machine learning	Social media	Covered features selection, imperfect data, etc.	--
A. Asbrino et al	Machine learning	Social media	- Reduced the number of unlabeled points, - Increasing performance	--
Matthew J. Denny et al	Machine learning	Twitter	Decreasing the risk for researchers	Understanding the researchers Opinions
Vivek Kumar et al	Machine learning	Social media and blogs	Language sentence extraction	Cannot understand the other language
Shichao zhang et al	Machine learning	Social media	- Produces concentrative patterns, - High quality performance	Data cleaning problems

PROPOSED WORK:

Data Preprocessing:

Data preprocessing is transforming the data into a basic form that make it easy to work. Data preprocessing is an important technique for processing the performance of Data Mining algorithms. In there are several preprocessing tools are available in DM. figure 1 shows data preprocessing techniques for Twitter data.

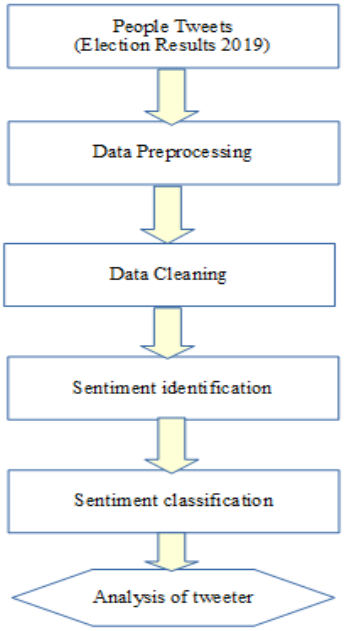

Figure 1: Twitter Data Preprocessing

Data Preparation:

Data preparation is a set of techniques that initialize the data properly to sense as input for containing DM algorithm. In this paper, the data is collected from Twitter data. The data is related to the Indian parliament election results.

Data Cleaning:

Data cleaning is an important technique for data preprocessing tool. It is a process of DM techniques. It removes the bad errors data and reduced unnecessary information of data. The missing of data are also included in data cleaning techniques. The presence of noise data may affect the intrinsic characteristic of a classification problem.

Stemming:

Stemming is a process of removing infectional words which is affixes, for example playing-play, studies-study. Stemming works on some particular language mainly English and Spanish.

Lemmatization:

Lemmatization takes the consideration of morphological analysis of the words. It reduces inflected words properly with the root words belongs to the sentences. It also called as lemma which is the set of words in dictionary form, citation form and canonical form.

Data reduction:

The data reduction represents the original data that reduced to obtain a set of techniques to one way or another way those data needed to the distinction of data preparation to approximately suit the input data of DM task.

Data Normalization:

The data normalization is used to analyzing measurement unit that is expressed in the attributes of measurement unit range and common scale.

a. Min-Max Normalization:

$$v' = \frac{\text{v} - \text{min}}{max - min}(\text{new\_max} - \text{new\_min}) + \text{new\_min}$$

b. Z-Score Normalization:

$$v' = \frac{\text{v}-\text{mean}}{\text{stand\_deva}}$$

Where v is the old feature value and v' is the new one.

Sentiment Classification:

In generally to calculate sentiment score for each tweet shown below:

Score = number of positive words - number of negative words

If score > 0 it is positive sentiment

Score < 0 it is negative sentiment

Score = 0 it is neutral sentiment

Algorithm:

Input		Training twitter data set t, Positive sentiment Pt, negative sentiment nt, neutral sentiment nut
Output		Sentiment Classifier C, Total tweets tw
Step	**1**	Import Twitter API
	2	Set the twitter authentication // create twitter account API
	3	Create the class and set the data cleaning method Create stemming and lemmatization function
	4	Try If sentiment > 0 : positive Else if sentiment < 0 : negative Else : neutral
	5	Main definition Call the twitter class function
	6	Find the Percentage of Twitter sentiment score //accuracy
	7	Print the twitter data

Table 2: Twitter data sentiment calculation

Total Twitter Words	1000
Positive Sentiment Percentage	32 %
Negative Sentiment Percentage	18 %
Neutral Sentiment Percentage	39 %
Total Accuracy Percentage	89 %

Table 2 shows the results of Twitter sentiment analysis percentage. This research work has collected 1000 twitter words in twitter website using the hashtag of #electionresults2019. The proposed algorithm easily to identify the people opinion and find the sentiment score of Twitter sentiment data.

CONCLUSION:

Data preprocessing is a process of DM algorithms. The proposed work is identifying the people sentiment in election results in 2019. This algorithm improves the accuracy of Twitter data. The noisy data and stop words are removed and given better results in data cleaning algorithms. This algorithm easily to classify the better results in data preprocessing techniques. In the future to improve the data quality and find the context words in sentiment. Moreover to improve the good accuracy of Twitter data.

REFERENCES:

[1] Jayaram Hariharakrishnan, Mohanavalli.S, Srividya, and Sundhara Kumar K.B "Survey of Pre-processing Techniques for Mining Big Data," IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017), Year: 2017

[2] Chen Min, Shiwen Mao, and Yunhao Liu. "Big data: a survey" Mobile Networks and Applications, PP: 171-209. 2014.

[3] Przemyslaw Grzegorzewski and Andrzej Kochanski "Data Preprocessing in Industrial Manufacturing" Springer Nature Switzerland AG, DOI: https://doi.org/10.1007/978-3-030-03201-2_3, Year: 2019

[4] STAMATIOS-AGGELOS N. ALEXANDROPOULOS, SOTIRIS B. KOTSIANTIS and MICHAEL N. VRAHATIS "Data preprocessing in predictive data mining" Cambridge University Press, Vol. 34, 1–33. DOI: 10.1017/S026988891800036X, Year: 2019

[5] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, Jose Manuel Benítez, and Francisco Herrera "Big data preprocessing: methods and Prospects" Big Data Analytics, DOI: 10.1186/s41044-016-0014-0, Pp: 1-9, Year: 2016

[6] A. Astorino, E. Gorgone, M. Gaudioso & D. Pallaschke "Data preprocessing in semi-supervised SVM classification" DOI: 10.1080/02331931003692557, VOL: 60:1-2, PP: 143-151, Year: 2017

[7] Matthew J. Denny and Arthur Spirling "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It" DOI: https://doi.org/10.1017/pan.2017.44, Year: 2017

[8] Vivek Kumar, Abhishek Verma, Namita Mittal and Sergey V. Gromov "Anatomy of Preprocessing of Big Data for Monolingual Corpora Paraphrase Extraction: Source Language Sentence Selection" Springer Nature Singapore Pte Ltd, DOI: https://doi.org/10.1007/978-981-13-1501-5_43, Year: 2019

[9] Shichao Zhang, Chengqi Zhang and Qiang Yang "Data preparation for data mining Data preparation for data mining," Applied Artificial Intelligence, DOI: 10.1080/713827180 VOL: 17:5-6, PP 375-381, Year: 2018

[10] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas "Data Preprocessing for Supervised Learning" INTERNATIONAL JOURNAL OF COMPUTER SCIENCE VOLUME 1, ISSN 1306-4428, Year: 2006