International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017), e-ISSN: 2455-2585
Volume 4, Issue 08, August-2018

DATA PRE-PROCESSING:REVIEW ON VARIOUS PRE-PROCESSING METHODS IN OPINION MINING

[1]Manisha Sharma, [2]Prof. Nirupama Tiwari

[1]Computer science & Engineering, Shriram college of engineering & management, Banmore, India
[2]Computer science & Engineering, Shriram college of engineering & management, Banmore, India

***Abstract*— Opinion mining is a different method right from mining classification to advance opinion mining. First, traditional techniques were discussed for solving the problem of opinion mining. It is study that approaches based on pre-processing method. Data preprocessing involve some steps with data compilation, data cleaning, session classification, user recognition and pathway completion. This paper present numerous data preprocessing technique in arrange to organize unprocessed data appropriate used for mining & analysis tasks.**

***Keywords*— Data Mining, Opinion Mining, Data pre-processing,pre-processing methods**

I. INTRODUCTION

Data mining is used to extract implicit and previously unknown information from data. DM is the procedure this gives an idea to draw in consideration of clients because of High accessibility of colossal measure of data and need to change over such data into useful information. [1] .The opinion mining is used world widely in real time applications. In simple terms, Opinion mining is a type of common technique used for identifying whether the public is interested or not in a product. Opinion mining aims to observe the people's opinions or sentiments and attitudes of the products services and product attributes. The Sentiments or user opinions expressed in textual format are obtained from different websites or mobile apps. For Example, the dataset level opinion mining captures the subjectivity and important sentiments expressed in a review dataset [2] Opinion mining is used to express or describe the opinions of users. Users are given positive and negative opinions using the services. These opinions are very helpful in making some purchase decisions. The accurate identification of correct opinion is very critical to identify within the large textual dataset and from unstructured datasets. It is a must to design an efficient method to identify the features from the unstructured datasets. The savvy needs the region of receiving the rating to know the positive or negative contents of the final products rating. opinion mining a evaluation tin be evaluated at 3 different levels- at article level, sentence level with feature level. At the point when audit is assessed at sentence level, at that point each sentence in a survey is arranged into either positive or negative. While include level or highlight based opinion mining gives synopsis which highlight of item is liked or hated by reviewer. [3]

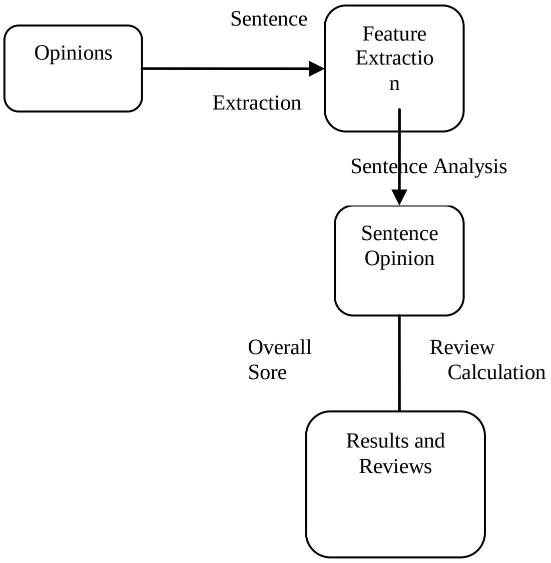

Figure 1: Opinion Mining Process

II. PRE-PROCESSING IN OPINION MINING

Data Preprocessing is a process to represent the data in format as per mining techniques. Data preprocessing comprises of interactive advances: data cleaning, data integration, data selection, data alteration among so on. [4]

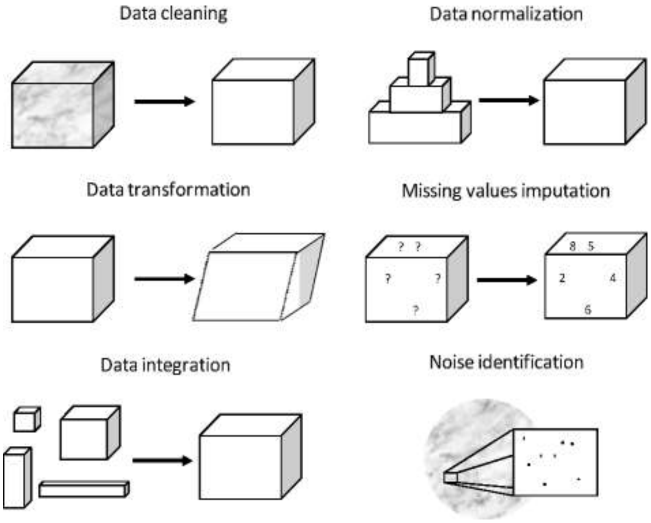

Fig 2: Data pre-processing method

A. Data cleaning

It is the demonstration of detecting & correcting (or evacuating) degenerate or off base records from a record set, table, or database. Utilized for the most part in databases, the term alludes to distinguishing fragmented, off base, mistaken, immaterial and so forth. Parts of the data and after that supplanting, changing or erasing this grimy data. Genuine data have a tendency to be incomplete, loud, and conflicting. Data cleaning schedules endeavor to fill in missing qualities, smooth out clamor while recognizing exceptions, and right irregularities in the data.

Missing Values: Filling in the missing qualities for the specific attribute.. it comprises of following techniques:

Ignore the tuple: This is generally done when the class label is missing. This strategy isn't exceptionally effective, except if the tuple contains a few characteristics with missing values.

Fill in the missing value manually: In general, this approach time-consuming and may not be practical given an expansive data set with numerous missing values.

Use a global constant to fill in the missing value: Supplant all missing attribute by a similar steady, for example, a name like "Unknown" . On the off chance that missing values are supplanted by, say, "Unknown," at that point the mining system may erroneously believe that they shape a interesting concept, since they all have an esteem in common—that of "Unknown." Hence, although this method is simple, it is not fool proof.

Use the attribute mean to fill in the missing value: For instance, assume that the normal wage of All Electronics customers is $56,000. Use this value to replace the missing value for income.

Use the attribute mean for all samples belonging to the same class as the given tuple: For example, Utilize the attribute mean for all examples having a place with an indistinguishable class from the given tuple: For instance, if classifying customers as indicated by credit hazard, supplant the missing value with the normal income incentive for clients in a similar acknowledge chance classification as that of the given tuple.

Use the most probable value to fill in the missing value: this might be resolved with regression, inference based instruments utilizing a Bayesian formalism, or decision tree acceptance.

a) Noisy Data: Noise is an random mistake or change in a deliberate variable. The data to evacuate the Noise utilize the accompanying data smoothing methods:[5]

Binning: Binning strategies smooth an arranged data esteem by counseling its "neighborhood," that is, the qualities around it. The arranged qualities are distributed into a no. of "buckets," or bins. Since binning techniques counsel the area of qualities, they perform neighborhood smoothing. In smoothing by container limits, the base and greatest qualities in a given canister are recognized as the receptacle limits. Each canister esteem is then supplanted by the nearest limit value. When all is said in done, the bigger the width, the more noteworthy the impact of the smoothing. Then again, receptacles might be equivalent width, where the interim scope of qualities in each bin is constant. Binning is likewise utilized as a discretization method.

Regression: Data can be smoothed by fitting the data to a limit, for instance, with relapse. Linear regression includes finding the "best" line to fit two attributes (or variables), with the goal that one attribute can be utilized to anticipate the other. Different linear regression is an expansion of linear regression, where in excess of two characteristics are included and the data are fit to a multidimensional surface.

Clustering: Anomalies might be distinguished by clustering, where comparative qualities are sorted out into groups, or "clusters."Instinctively, values that fall outside of the arrangement of clusters might be thought about anomaly analysis.

B. Data integration

Data integration, which joins data from various sources into a lucid data store, as in data warehousing. A property might be repetitive on the off chance that it can be "derived" from another attribute or set of attributes. Irregularities in attribute or measurement naming can likewise cause redundancies in the resulting data set. A few redundancies can be recognized by connection examination. it includes joining data residing in various sources and giving clients a unified together perspective of these data. This process becomes significant in a variety of situations both commercial and scientific.

C. Data selection

Precedes the real routine with regards to data collection.This definition recognizes data selection from particular data detailing and intuitive/dynamic data determination. The way toward choosing reasonable data for an research venture can affect data integrity.

Data transformation

It changes over data from a source data design into goal data. Data transformation can be isolated into two stages: data mapping and code age. In data change, the data are changed or consolidated into shapes appropriate for mining.[6]
Smoothing: This attempts to expel noise from the data. Such strategies incorporate binning, regression, & clustering.
Aggregation :where outline or aggregation activities are connected to the data. For example, the day by day deals data might be aggregated in order to process month to month and yearly aggregate sums. This progression is regularly utilized as a part of developing an data cube for examination of the data at various granularities.
Generalization: The data, where low-level or "primitive" data are supplanted by higher-level amount ideas using idea hierarchies. For instance, unmitigated attributes, similar to road, can be summed up to higher-level ideas, similar to city or nation. Essentially, values for numerical attributes, similar to age, might be mapped to higher-level ideas, similar to youth, moderately aged, and senior.
Normalization: where the property data are scaled in order to fall inside a little indicated go, such as1:0 to 1:0, or 0:0 to 1:0.Attribute development , where new attributes are built and included from the given arrangement of attributes to enable the mining to process. A attribute be normalized by scaling its qualities so they fall inside a little indicated go, for example, 0.0 to 1.0.Normalization is especially valuable for classification algo including NN, or distance estimations, for example, closest neighbor order and clustering . If utilizing the NN back proliferation algo for classification mining, normalizing the I/P esteems for each characteristic estimated in the preparation tuples will enable speed to up the learning stage.

Data Reduction

Data reduction systems can be connected to acquire a lessened portrayal of the data collection that is considerably littler in volume, yet nearly keeps up the integrity of the original data. This step looks to reduction of transformed data by extracting important features for mining technique. Strategies for data reduction include the following:
Data cube aggregation, where aggregation tasks are connected to the data in the development of an data cube.
Attribute subset selection: where irrelevant, weakly relevant, or repetitive attributes or measurements might be identified and evacuated..
Numerosity reduction: where the data are supplanted or evaluated by elective, littler data representations , for example, parametric models or nonparametric strategies, for example, clustering, inspecting, and the utilization of histograms.
Discretization and concept hierarchy generation: where raw data esteems for attributes are supplanted by ranges or higher reasonable levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies. Discretization and idea hierarchy generation are great devices for DM, in that they permit the mining of data at various levels of deliberation. [7]

III. LITERATURE SURVEY

Mika V. Mäntylä et al [2018] in late years, slant examination has moved from breaking down online item surveys to web based life writings from Twitter and Face-book. Numerous themes past item audits like securities markets, elections, debacles, medicine, software engg. and digital tormenting broaden the usage of slant analysis1.[8]

Deepak Soni et al [2017] In this paper, a simple logistic data model is applied on the data collected from twitter, face-book and other media during the time of elections and opinion can be analyzed based classification of data on bases of high and low popularity parties during elections. The results of our data model are very stimulating and good for the analysis. [9]

Shokoufeh Salem Minabet al [2015] the explanation behind this paper is to show the past works online examination of assumption on Twitter. Social media like Twitter create space to explain the thoughts and opinions on various topics and different events, millions of users can share their ideas in this Micro blog. Therefore Twitter is changed over as a main source to investigation of data; settle on a choice and an examination of supposition. There is a mean in the most extreme piece of the compositions, yet it is more fundamental to offer methods to gaining proper measuring and enhanced use of information for envisioning supposition. In like manner twitter information process after the stream illustrates. In this process, data were arrived base at rapid to destination in form of result; data mining algorithm should be capable to find user feeling in immediate time under limited space and time constraints. [10]

Lavanya T [2016] et al. It is proposed to design a algo which expels conclusion targets and opinion words using word course of action show for online reviews removed from Twitter. An opinion target is represented as the topic about which users shows their opinions. A opinion words are portray as the words that are utilized to speak to client's opinions purpose of the wander is to choose the examinations of blog or review creator with respect to some topic or the general pertinent furthest point of online overviews using word plan show. The aim of this project is to design an algorithm that predict all kind of opinion words and opinion targets for analyzing the market status of a product by mining client surveys posted on internetworking webpage to be specific the Twitter[11].
MD. Azza F. Yatim et al. [2016] this technique utilizes vocabulary for recognizing supposition of specific question inside related data set. This paper exclusively focuses on building the dictionary for the technique. By concentrating on Indonesian politic issues, we make another corpus way to deal with construct a relevant dictionary which utilizes news articles as corpora. We decide the underlying or essential seed words and have it examined by space specialists for our investigation.. We utilize this finding to assess and enhance our strategy as we proceed with the exploration to acquire more important result [12].

Jorge A. Balazset al [2015] In this paper we presented a short survey of the most popular Opinion Mining techniques, de_ned the Information Fusion _eld, proposed a simple framework for guiding the fusion process in an Opinion Mining system and reviewed some of the studies that have successfully implemented Information Fusion techniques in the Opinion Mining setting. In reality, the fate of Opinion Mining depends on making better and more profound wellsprings of information, which can be accomplished by melding effectively existing learning bases such as ontologies and lexicons. Nevertheless, few studies have done so by explicitly applying well-established techniques. Truth be told, studies in which creators meld di_erent lexical assets or strategies without following any standard methodology are the most widely recognized. [13]

Harpreet Kaur et al [2017] this paper presents a survey of sentiment analysis and classification algorithms. This survey concludes that sentiment classification is still an open field for research. There is a lot of scope for algorithms in it. SVM and naïve bayes are most popular algorithms for sentiment classification. Sentiment analysis of tweets is very popular. Datasets from sites like Amazon, IMDB, flip-kart are widely used for sentiment analysis. Deeper analysis is required in case of social networking sites. In many cases, context consideration is very important. Therefore more research is required in this field. [14]

IV. CONCLUSION

Data mining is a technology using which we can extract useful information from data. This paper shows a brief explanation on pre-processing techniques. It begins with pre-processing procedures which incorporates detailed depiction of different data cleaning approaches, and disregards the noisy data, imbalanced data handing and dimensionality reduction.

References

[1] Kalyani et al., International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X ,Volume 2, Issue 10, October 2012.

[2] Xiaohui Yu, Member, IEEE, Yang Liu, Member, IEEE, Jimmy Xiangji Huang, Member, IEEE, and Aijun An, Member, IEEE," Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 4, APRIL 2012.

[3] Fuji Ren, Senior Member, IEEE, and Ye Wu," Predicting User-Topic Opinions in Twitter with Social and Topical Context" IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL.4,NO.4,OCTOBER-DECEMBER 2013.

[4] Jing Lu et al., "Timeline and Episode - Structured clinical data: Pre-processing for data mining and analytics", ICDE 2016 IEEE Workshops.

[5] Soukup, T., & Davidson, I. (2002). Visual data mining: Techniques and tools for data visualization and mining, Wiley.

[6] J. Wang and G. Karypis. On anciently summarizing categorical databases. Knowledge and Information Systems, 9(1):19{37, January 2006.

[7] Jaiwei Han and Micheline Kamber, "Data Mining Concepts and Techniques", Second Edition, ISBN 13:978-1-55860-901-3.

[8] Mika V. Mäntylä, Daniel Graziotin, Miikka Kuutila, "The Evolution of Sentiment Analysis - A Review of Research Topics, Venues, and Top Cited Papers", Volume 27,February 2018.

[9] Deepak Soni, Mayank Sharma, Sunil Kumar Khatri, "Political Opinion Mining Using E-social Network Data", 978-1-5386-0514-1/17/$31.00 ©2017 IEEE.

[10] Shokoufeh Salem Minab, Mehrdad Jalali and Mohammad Hossein Moattar "Online Analysis of Sentiment on Twitter" 2015 IEEE.

[11] Lavanya. T, Miraclin Joyce Pamila. J. C, Veningston. K "Online Review Analytics using Word Alignment Model on Twitter Data"2016 IEEE.

[12] MD. Azza F. Yatim, Yulistiyan Wardhana, Ahmad Kamal “A Corpus-Based Lexicon Building in Indonesian Political Context Through Indonesian Online News Media” 978-1-5090-4629-4/16/$31.00@ 2016 IEEE.

[13] Jorge A. Balazs and Juan D. Vel_asquez[2015],” Opinion Mining and Information Fusion: A Survey”, June 4, 2015.

[14] Saif, Hassan, et al. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management 52: 5-19,2016.