International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Text Analytics on Un-structured Text Data using ANN

ShivKumar Goel[1], Anil Kumar Bhandi[2]

[1]Professor, Deputy Head of Department, MCA, VESIT, University of Mumbai, Mumbai, India-400074

[2]Department of MCA, VESIT, University of Mumbai, Mumbai, India-400074

Abstract: ***In today's world of fast growing technology the emerging word Big Data has acquired the attention of major organizations for their business. Most of the data is in unstructured format; interpreting such complex and informal data to take decision is an emerging challenge. The proposed paper is about extracting the useful data and to overcome the major issues in text-usually human errors. There are various human errors and colloquial which reduce the quality of text thus makes the text complex to analyse and weakening the decision making tool. The set of patterns and rules to search the data is limited and restricted whereas the errors made manually can be corrected by methodologies of soft computing especially Artificial Neural network (ANN). The proposed system takes the un-structured text data and using the Natural language processing rectifiesthe error and corrects it to get the maximum quality in text. The Analytical tools process the text to get the complete report and help in decision making. The neural network plays an important role by learning the errors and developing a training data for the future analysis on new unstructured text.***

Keywords: Artificial neural network (ANN), supervised learning, unsupervised learning, natural language processing (NLP), HADOOP, Big data, Text mining. Online transaction processing (OLTP), Reinforcement learning, SEO, search engine.

1. Introduction

Un-structured text data not just comprises of text and tables from the organization but also the huge bulk of Internet, PDF's, word files, chats, E-mails, social networking sites, E-commerce websites and many more. Various techniques can be easily applied on text that is structured in form of tables or specific format. The Analytics operation can be easily performed on structured data and a quite perfect outcome of the data can be analysed. The structured data mostly comes from OLTP of the organization easily converted in decision making. Major text data is un-structured that comes from the chats, mails, review on E-commerce website etc. These kind of raw data aredifficult to process as the format of the data is not specific and various virtual noise may effect it. The segregation of these data are carried out only through the set of rules and patterns defined in the analytics tool.

2. Literature Review

The text analytics process is carried out by gathering the text data from various domain and then this data is pre-processed and thusundergoes the analytical tool using various text mining techniques to extract the consistent data. This data is then used in generation of reports or decision making.

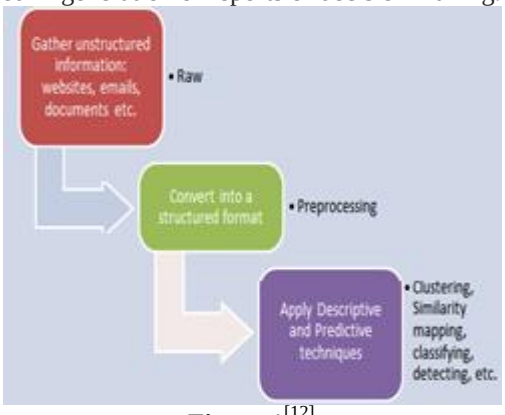

Figure 1[12]

The subtasks andcomponents of a larger text-analytics effort typically include:

- Information retrieval or identification of a corpus is a preparatory step: collecting or identifying a set of textual materials, on the Web or held in a file system, database, or content corpus manager for analysis.
- Although some text analytics systems apply exclusively advanced statistical methods, many others apply more extensive natural language processing, such as part, syntactic parsing, and other types of linguistic analysis.
- Named entity recognition is the use of gazetteers or statistical techniques to identify named text features: people, organizations, place names, stock ticker symbols, certain abbreviations, and so on. Disambiguation - the use of contextual clues - may be required to decide where, for instance, "Ford" can refer to a former U.S. president, a vehicle manufacturer, a movie star, a river crossing, or some other entity.
- Recognition of Pattern Identified Entities: Features such as telephone numbers, e-mail addresses, and quantities (with units) can be discerned via regular expression or other pattern matches.
- Co reference: identification of noun phrases and other terms that refer to the same object.
- Relationship, fact, and event Extraction: identification of associations among entities and other information in text
- Sentiment analysis involves discerning subjective (as opposed to factual) material and extracting various forms of attitudinal information: sentiment, opinion, mood, and emotion. Text analytics techniques are helpful in analysing, sentiment at the entity, concept, or topic level and in distinguishing opinion holder and opinion object.
- Quantitative text analysis is a set of techniques stemming from the social sciences where either a human judge or a computer extracts semantic or grammatical relationships between words in order to find out the meaning or stylistic patterns of, usually, a casual personal text for the purpose of psychological profiling etc.

The text data from various social networking site, E-commerce, mails, chats etc. is important form an organization's perspective; as it contains the product review, various information about the brand and detailed overview. Most organization consider this raw data and perform text analytics on it, but the major issues faced by them are language barrier, grammatical errors, colloquialetc. where rules and pattern of text are used to extract data statistics.

The analytical tools mostly used to filter the data is based on [11]:

- Knowledge based – set of rules defined mainly used in enterprise model
- Statistical data algorithm – not transparent uses common pattern specification language (CPSL – grammar based) or rule based information extraction.

Due to human error major part of the data can be unprocessed or discarded. The hidden pattern of information may be lost or undiscovered. The majority of the language errors can be eliminated by NLP. Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

3. Methodology

Most of the data come from social networking sites, E-commerce, mails etc. where the end user express their views about a particular topic or review the product usage and many more aspects. Organization process such data but most of time user use slang or maybe there are grammatical mistakes and sometimes data or text with noise i.e. text having no meaning. The un-structured data is pre-processed and used to find the information related to semantics, based on rule pattern. Our proposed system helps to eliminate the text noise or human error by using the NLP.

NLP process the unstructured text and provides a detailed description of error and the relative suggestion the statement should be grammatically correct. NLP mostly can take a firm decision based on such grammatical errors but the major errors in the text created by humans can only be corrected by human initially and later the proposed system-artificial neural network could play a crucial role in rectifying similar errors

Few human errors can be:

- Colloquial/Abbreviation
- Sentence breaking
- Wrong formation of sentence's
- Self-generated short form
- Incomplete statements.
- Unusual word.

There are many such errors made,while chatting and mailing in an organization. This communication contains lots of information but the quality of the text to process and analyse is inadequate. The software itself can't decide what the statement's concludes.

3.1 Functionality

The raw text is used in for Analytics. The first step is pre-processing where the huge text data is classified in to chunks based on the completion of the paragraph or the completion of the text. This chunks of data undergo the next phase i.e. NLP and the various errors are identified. Since most of the errors are human generated it can be corrected by human understanding by considering the option provided and taking the correct decision for a particular statement.The similar errors can be combined and the best solution for it can be taken based on the count, so that the data consistency, correctness and quality is improved. The text quality helps the statistical data algorithm to perform an analysis. The correction is done initially by the active user based on the suggestion provided by the NLP. The active user can go by the suggestion of NLP or can suggest the correct meaning of the word.

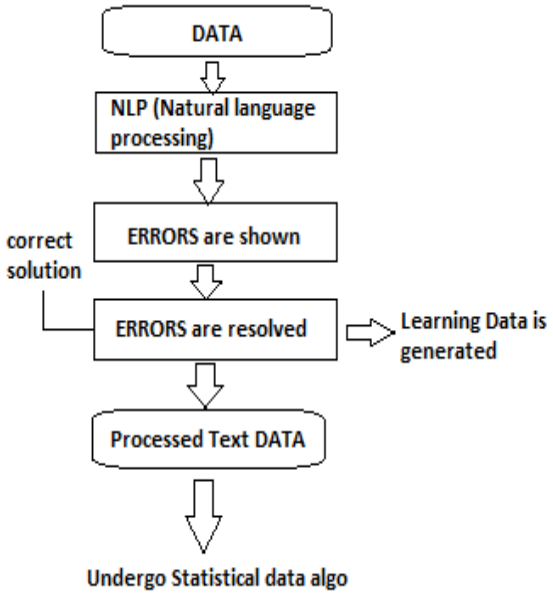

Figure 2: Filtering of raw text

Neural network and NLP are used to filter the quality of text**.** If the solution is specific the system takes the solution and processes the data and saves it in the learning data and act as supervised learning. If the solution is multivalued the neural network takes the solution and perform the reinforcement learning using the NLP and the decision is taken based on the better solution. Assumption, if an organization is looking for the product review in term of good, better, best, awesome or something not good, worst etc. and the data contains review like gud (good), grt(great), awsm(awosome), f9(fine) these sentences are not considered mostly as they do not follow the general syntax and semantics. This may affect the analytical tool in making the decision, leads to wrong statistics. The neural network takes the input for the first layer of nodes process to it by adjusting the weights and provides output.

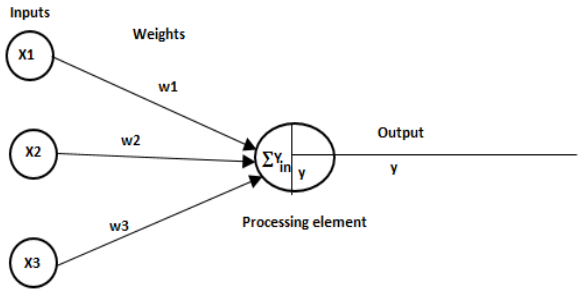

Figure 3: Neural network processing [10]

In the proposed system the inputs are the errors and weights are the suitable solution needed to be implemented on the sentences to make it error free. The processing element is the NLP that processes the error and changes the data throughout the chunk of data. The Neurons undergo simulation of various aspects of error in the data and learns the situation that need to be applied on the errors. The processing is done only by NLP with the various condition and the actual output is the correct statement with maximum correction. The artificial neural network process the errors efficiently and rectifies the raw text to valuable information.

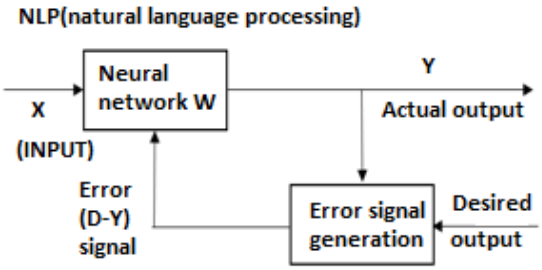

Figure 4: Supervised learning [10]

A multivalued solution for a problem can also be considered where the similar errors are taken and the suggested solution is set as the bias which are applied on the errors to get the suitable statement without changing the meaning and the weights. For this NLPbased suitable solution is used in rectifying errors.The outer layer of neurons process the node with each bias set and the NLP takes the decision adapting to the suitable solution for each error and corrects it. This method of correcting is carried out under the process of reinforcement learning using unsupervised learning.

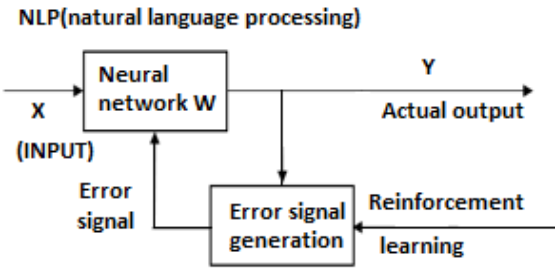

Figure 5: Reinforcement learning (unsupervised learning) [10]

Reinforcement learning is similar to the supervised learning. In supervised learning, the correct target output values are known for each input pattern. In some cases, less information might be available, or the NLP can frame the statement exactly but the solutions in the learning data may be 50% correct to get the actual output or so thus the critic information is available not the exact information. The Neural network process the error input with various learning data and try to get the correct text which is the actual output. The learning based on the critic information is called reinforcement learning and the feedback is called reinforcement signal.

The reinforcement learning sub-discipline of unsupervised learning, is used mainly when the solution is not accurate or the user lets the system to take correct solution from the training data and produce quality information. Most often the error are identified by the NLP and suggested with better solution, but if the solution is not appropriate then the active user can input the actual solution. The user can also provide multivalued solution to the system. The system initially works with lot of errors, the neural network gets trained with such errors and develop the training data. As the training data increases the network uses the training data usually onsimilar problems and further applies reinforcement learning to produce thequality in text.

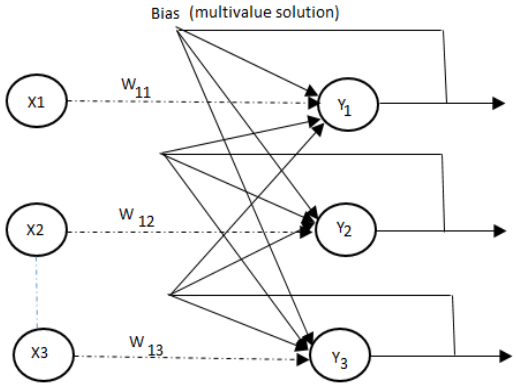

Figure 6: Single layer recurrent network [10]

The similar errors are taken as input to the network (X_1, $X_2 \ldots X_i$) where the weights (W_{11}, $W_{12} \ldots W_{ij}$) are the suggested value by the NLP and the bias is the various multivalued solution to be applied on the errors of similar type. The neurons (Y_1, $Y_2 \ldots Y_j$) are the network nodes that work as NLP provides output if the text is not filtered or corrected then error is generated and the statement is again put forth as errors; completely with description to active user to take decision by considering a detailed look through and provide a better solution. The error are feed as input to the input layer, the errors that are similar are being processed if the error are just single in count a single neuron itself can do the processing and generate the expected text, as the count increases layer of input to the network gets complex. The network is single layer recurrent network where the errors are processed of similar type the NLP suggestion is used as the initially to generate the text, if the user have suggested a single value then it takes the user's solution and gets the desired text. If there are multivalued then each value is used in recurrent network and the accuracy of the statement is evaluated and the best value is used to filter the text in the complete chunk. A similar process is carried out for the errors encountered and accepted as multivalued input; by setting each as bias and further evaluated to get the appropriate statement. The NLP plays an important role in evaluating the value to statement and getting the rich text. Even after completion of learning data, if the reinforcement

learning is not successful then the system takes a suitable input for the individual error and corrects as per user's suggestion.

This process of initial phase provides a rich text that is error free and all the human errors are removed to get the maximum accurate information for applying statistical data algorithm. The process of filtering the datagets accurate and perfect as the training data increases, thus the neural network behaves like a human brain and takes the decision as a human brain is capable of taking. The major issues of social networking and E-commerce websites data is eliminated.

Application of this step is, the text data can be classified such as weather data, product review data etc. The NLP while processing the data chunks it can go through the various conditions and process the text to get the data based on various domain. The data can further undergo the common pattern specification or rules based information extraction to extract the information needed for analytics. Association among the entity can be one of the relation to extract the relation data such as any user name, address, state, pin code, contact numbers etc.

3.2 Layout of Proposed System

The working of the system is functional and crucial system is the network. The flow of our system is as follows:

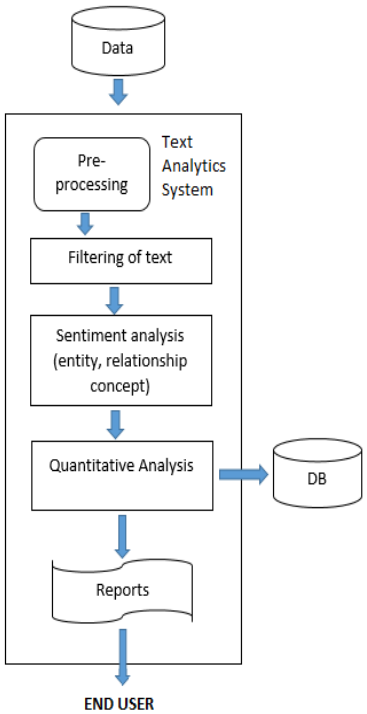

Figure 7: Work flow of the system

The next step in the proposed system is the main mining of the exact data where the sentiment analysis takes place. The sentiment analysis uses the statistical data algorithm i.e. Rule based information retrieval and common pattern specification language where set of rules are being written and the system search that pattern to get the data. The sentiment analysis deals with the entity relationship concept where the user details or the relation information of the particular entity is searched. The pattern are formed based on the language grammar base i.e. if a word is to be searched from the processed text the various form of it is searched such as its noun, adverb, verb, adjective. The word tends to be similar synonym but the way of using in the sentences differs so many patterns and rules based syntax are written in order to process the text.

The processing of the text by a local system is difficult so a big data tool such as HADOOP, can be used to process the text and get the results faster. The rule based syntax can be written in the mapper and the mapper does the work of searching the text. As the mapper completes its work;the reducer performs the main task of Quantitative analysis. The Reducer combines the count of the words that was given as input to be searched. The system results with a detailed layout of the words in various form and their count.

3.3 System Forms

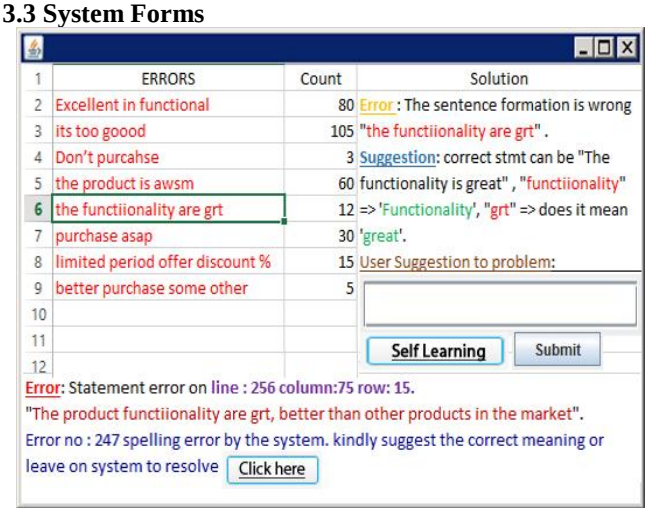

Figure 8: Error report

The system provides a way to insert the raw data. The data can be any unstructured text form. The data is divided into clusters and each data cluster undergo the operation. The filtering of the text takes place where the human error is shown and the NLP suggest the solution in the form of row/tab. The user can go with NLP suggestion by clicking the self-learning or click here option. The errors undergoneural network operations or the active user can provide a solution for it. As training data increases the system performs the reinforcement learning and the filtration of text; else the errors are shown in the window fig (8).

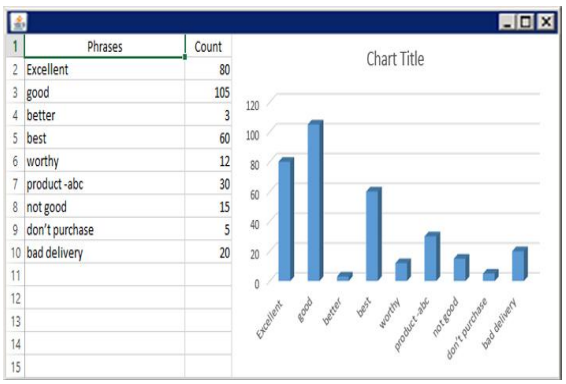

Figure 9: Detailed report

After the completion of the error evaluation the algorithm performs the analysis of the entity using the HADOOP environment. The Quantitative analysis is carried out and the count of the words are shown in form of report. The reports could be used by the organizationin decision making. The text analytics of the system is very accurate; the system at the initial stageremoves the error with some human intervention and later the neural network behaves like an self-learning system; as most of the errors are evaluated by system itself without waiting for the user to give suggestion as the system itself has the scenario of the previous learning.

4. Application

The proposed system concept can be used in various application. The learning concept in our system can be added to search engines, search engine optimization, social networking applications, E-commerce brand/product reviews and many more. In search engines the user queries the particular text, to get the related information, the system can be implemented similarly where the data is being filtered from the various pages and related links can be shown as well the network maintains its learning data for similar queries. The search engine optimization can also be done as the text that is being filtered or searched more often the URL of the pages can be kept in the memory and when new user searches it once again the user can get the best content related to the queries. The system well suitable for social networking trends and the various customer review about the upcoming event can be known easily by just using the social network data rest all the reports can be modified to some extent as per the requirements.

5. Conclusion

The technology is getting faster with new trends, the amount of data is increasing by leaps and bound from mobiles as compared to the personals computers. The social media, e-commerce is a stream of getting the huge amount of data and filtering the data for complete use is challenging, the proposed system get an accurate rich text for analytics. The other unstructured data such as image, videos, sound etc. filtering and performing analytics is more challenging as image processing and various pattern matching techniques would help to some extent. The various pattern matching techniques used in text mining are efficient in getting the text data and relation among them. The hidden pattern of the text information can be used by the organization in taking a firm decision for the growth and achievement of the goal as well as to know the end user want from the organization.

References

[1] Dr.Goutam Chakraborty “Analysis of Unstructured Data: Applications of Text Analytics and Sentiment Mining”, Paper 1288-2014
[2] “Integrated Visual Analytics Tool for Heterogeneous Text Data”IEEE Symposium on Visual Analytics Science and Technology October 24 - 29, Salt Lake City, Utah, USA 978-1-4244-9487-3/10/$26.00 ©2010 IEEE
[3] Alex Endert, Patrick Fiaux, Chris North” Semantic Interaction for Visual Text Analytics”
[4] Vishal Gupta, Gurpreet S. Lehal “A Survey of Text Mining techniques and applications, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 1, NO. 1, AUGUST 2009
[5] Text Mining Techniques- A Survey “International Journal of Advanced Research in Computer Science and Software Engineering “Volume 2, Issue 4, April 2012 ISSN: 2277 128X
[6] Big Data and Text Analytics: Solving the Missing “Big Language” Link, Forrester Research, Inc., 2013 Customer Experience Predictions, January 2013
[7] “Visual Analytics of Text Streams Through Multiple Dynamic Frequency Matrices”, IEEE Symposium on Visual Analytics Science and Technology 2014, November 9-14, Paris, France.
[8] “Integrated Visual Analytics Tool for Heterogeneous Text Data”, IEEE Symposium on Visual Analytics Science and Technology 2014, November 9-14, Paris, France.
[9] “Word Cloud Explorer: Text Analytics based on Word Clouds”, 2014 47th Hawaii International Conference on System Science. 978-1-4799-2504-9/14 $31.00 © 2014 IEEE DOI 10.1109/HICSS.2014.231.
[10] Book - Principles of Soft computing , Author - Sivanandam and Deepa
[11] http://www.ibm.com/software/data/bigdata/
[12] http://www.simafore.com/blog/bid/116340/Keyword-clustering-using-web-mining-and-text-mining-with-RapidMiner -posted by Bala Deshpande.

Author Profile

Anil Kumar Bhandi: Student of MCA from Vivekanand Education Society’s Institute of Technology, Chembur (University of Mumbai). Bachelor’s Degree in BScIT (Information Technology) from S.I.W.S College, Wadala (University of Mumbai).

Prof. Shivkumar Goel: Deputy Head of Department MCA inVivekanand Education Society’s Institute of Technology, Chembur (University of Mumbai).