ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

Formation of Smart Sentiment Analysis Technique for Big Data

Manisha Shinde-Pawar

Assistant Professor, Dept. of Management, IMRDA, SANGLI, Bharati Vidyapeeth Univerisity, Pune, India

ABSTRACT: Many of the top digital e-textbook companies employ big data in the form of analytics to not only measure customers buying habits, but also to provide the organizations with measurable data. Analytics are more important than just clicking on a buy button. Analyzing the voluminous data at an instant of time in memory to take right decisions is great challenge. To avoid such situations the basic need is to study sentiments while taking decisions. Here data analytics can help to analyze such big data. This has given rise a thirst for carrying out the study on sentiment analytics, big data and use of some smart algorithm to discover correct sentiments or opinions from unstructured big data.

The approach uses natural language processing techniques of Artificial Neural Network to extract features of interest from textual data retrieved from a micro blogging platform in real-time and, hence, generate appropriate executable code for the Decision Science and get predetermined means of social communication. So by enriching semantic knowledge bases using Fuzzy Logic (for fitness approximation) for Opinion Mining in Big Data Applications with predetermined means, suggested user action decisions can be improved.

KEYWORDS: Artificial Neural Network, Big Data, Decision Science, Fuzzy Logic, Opinion Mining

I. INTRODUCTION

A. How new smart sentiment analysis technique will help to solve problem

It's about combining and analyzing data so you can take the right action, at the right time, and at the right place.

What is sentiment analysis?
Sentiment analysis(also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials.

Sentiment Analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

A basic task in sentiment analysis is classifying the polarityof a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

B. Subtasks
A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

C. Methods and features

Existing approaches to sentiment analysis can be grouped into four main categories: keyword spotting, lexical affinity, statistical methods, and concept-level techniques. Keyword spotting classifies text by affect categories based on the presence of unambiguous affect words such as happy, sad, afraid, and bored. Lexical affinity not only detects obvious affect words, it also assigns arbitrary words a probable "affinity" to particular emotions. Statistical methods leverage on elements from machine learning such as latent semantic analysis, support vector machines, "bag of words" and Semantic Orientation — Point wise Mutual Information. More sophisticated methods try to detect the holder of a sentiment (i.e. the person who maintains that affective state) and the target (i.e. the entity about which the affect is felt). To mine the opinion in context and get the feature which has been opinionated, the grammatical relationships of words are used. Grammatical dependency relations are obtained by deep parsing of the text. Unlike purely syntactical techniques, concept-level approaches leverage on elements from knowledge representation such as ontologies and semantic networks and, hence, are also able to detect semantics that are expressed in a subtle manner, e.g., through the analysis of concepts that do not explicitly convey relevant information, but which are implicitly linked to other concepts that do so.

A human analysis component is required in sentiment analysis, as automated systems are not able to analyze historical tendencies of the individual commenter, or the platform and are often classified incorrectly in their expressed sentiment. Automation impacts approximately 23% of comments that are correctly classified by humans.
Sometimes, the structure of sentiments and topics is fairly complex. Also, the problem of sentiment analysis is non-monotonic in respect to sentence extension and stop-word substitution (compare *THEY would not let my dog stay in this hotel* vs *I would not let my dog stay in this hotel*). To address this issue a number of rule-based and reasoning-based approaches have been applied to sentiment analysis, including Defeasible Logic Programming. Also, there is a number of tree traversal rules applied to syntactic parse tree to extract the topicality of sentiment in open domain setting.

D. Evaluation

The accuracy of a sentiment analysis system is, in principle, how well it agrees with human judgments. This is usually measured by precision and recall. However, according to research human raters typically agree 79% of the time (Inter-rater reliability). Thus, a 70% accurate program is doing nearly as well as humans, even though such accuracy may not sound impressive. If a program were "right" 100% of the time, humans would still disagree with it about 20% of the time, since they disagree that much about *any* answer. More sophisticated measures can be applied, but evaluation of sentiment analysis systems remains a complex matter. For sentiment analysis tasks returning a scale rather than a binary judgment, correlation is a better measure than precision because it takes into account how close the predicted value is to the target value.

E. Sentiment analysis and Web 2.0

The rise of social media such as blogs and social networks has fueled interest in sentiment analysis. With the proliferation of reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for businesses looking to market their products, identify new opportunities and manage their reputations. As businesses look to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and actioning it appropriately, many are now looking to the field of sentiment analysis. If web 2.0 was all about democratizing publishing, then the next stage of the web may well be based on democratizing data mining of all the content that is getting published.

The problem is that most sentiment analysis algorithms use simple terms to express sentiment about a product or service. However, cultural factors, linguistic nuances and differing contexts make it extremely difficult to turn a string of written text into a simple pro or con sentiment. The fact that humans often disagree on the sentiment of text illustrates how big a task it is for computers to get this right. The shorter the string of text, the harder it become.

F. Natural language processing

Natural language processinggives machines the ability to read and understand the languages that humans speak. A sufficiently powerful natural language processing system would enable natural language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval (or text mining) and machine translation.

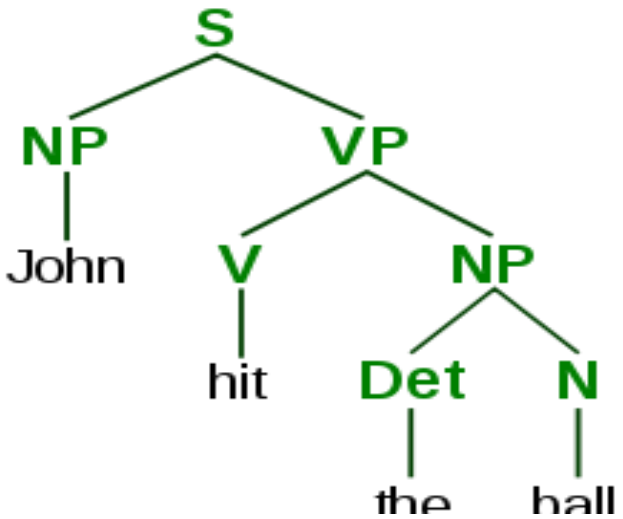
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

Figure No.1 Natural Language Processing (NLP)



As shown in Figure No.1, a parse tree represents the syntactic structure of a sentence according to some formal grammar.A precise set of evaluation criteria, which includes mainly evaluation data and evaluation metrics, enables several teams to compare their solutions to a given NLP problem. Here the Figure No. 1 represents parse tree with text as in memory data structure of variety of collection tokens with ability of quick text analysis with frequency, collocation, similarity, and simple regex-based searching. A `TextCollection` is a grouping of `Text` instances that allows you to do corpus-wide calculations (frequency, inverse document frequency etc).

G. BIG DATA

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, connectomics, complex physics simulations,and biological and environmental research. The limitations also affect Internet search, finance and business informatics. The easiest definition of big data as given by Sam Madden in white paper "From databases to Big Data" is stated as "Data that is too big, too fast, or too hard for existing tools to process".

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. Big data is distributed data. This means the data is so massive it cannot be stored or processed by a single node. It's been proven by Google, Amazon, Facebook, and others that the way to scale fast and affordably is to use commodity hardware to distribute the storage and processing of our massive data streams across several nodes, adding and removing nodes as needed.

The data is said to be Big Data if it is characterized by Volume, Velocity and Variety. In addition to these the purpose of data is to create value, and the complexity which increases because of the degree of interconnectedness and interdependence in data.

The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte's (2.5×10^{18}) of data were created. The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Big data requires exceptional technologies to efficiently process large quantities of data within tolerable elapsed times. A 2011 McKinsey report suggests suitable technologies include A/B testing, crowdsourcing, data fusion and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

integration, genetic algorithms, machine learning, natural language processing, signal processing, simulation, time series analysis and visualization.

Issues & Challenges with Big Data

As above we have discussed sources above, according to Bill Frank the data in big data can be categorized as: Automatically generated by a machine, typically an entirely new source of data (eg. blogs), not designed to be friendly (e.g. Text streams), may not have much values (need to focus on the important part).
From the above categorization author has found the big data as

- Structured-Most traditional data sources
- Semi-structured-Many sources of big data
- Unstructured-Video data, audio data

According to Stephen Kaisleret. al. the data stored with machine plays very important role in decision making and knowledge discovery. A major challenge for IT researchers and practitioners is that growth rate is fast exceeding our ability to both:
(1)Design appropriate systems to handle the data effectively
(2) Analyze it to extract relevant meaning for decision making.
According to Michael Cooper & Peter Mell from NIST the issues associated with Big Data can be given as-

• Taxonomies, ontologies, schemas, workflow
• Perspectives – backgrounds, use cases
• Bits – raw data formats and storage methods
• Cycles – algorithms and analysis
• Screws – infrastructure to support Big Data

II. LITERATURE REVIEW

Till today different sentiment analysis techniques are being implemented with different aspects of evaluation for big data opinion analysis.

Table No.1 comparative Sentiment Analysis Technique

Sr. No.	Sentiment Analysis Technique	Takes into account
1	Document level sentiment analysis	Classifying the whole document as positive or negative
2	Supervised learning techniques Unsupervised learning techniques.	'terms and their frequency', 'parts of speech', 'sentiment words and phrases', 'sentiment shifters' Use of fixed syntactic patterns that occur in an opinion.
3	Sentence level Sentiment Analysis	Associated with a phrase or *sentence*
4	Aspect Based Sentiment Analysis	sentiment on entities and/or aspect of those entities
5	Oracle Advanced Analytics	database into a comprehensive advanced analytics

The above table no. 1 shows comparative analysis of different sentient analysis techniques and what it takes into account for implementation base. Polarity, subjective detection and opinion identification all are very important things in this kind of sentiment analysis. Document level sentiment analysis classifies the whole document as positive and negative statement documents. Supervised learning verifies terms and frequency, 'parts of speech', 'sentiment words and phrases', 'sentiment shifters'. Unsupervised learning technique uses fixed syntactic patterns that occur in an

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

opinion. Sentence level sentiment analysis is associated with phrase or sentence evaluation. Aspect based sentiment analysis evaluates sentiment and aspect of entities. And Oracle Advanced analytics provides database into comprehensive advanced analytics.

III. RESEARCH GAP

The creation of analytics and the consumption of analytics are two different things. The real challenge is transforming the people and the processes to analyze unstructured big data to extract relevant meaning. Thus researcher pointed a problem to design and develop sentiment analysis techniques for unstructured big data for translating analytics into good decisions.

IV. STATEMENT OF THE PROBLEM

As information technology system become less monolithic and more distributed, real-time big data analysis will become less exotic and more common place. At that point, the focus will shift from data science to next logical frontier: decision science.
Researches would like to focus big data for the study to design an efficient sentiment analytic technique for unstructured big data. Such Sentiment analysis is very useful to identify and predict current and future trends, product reviews, people opinion for social issues, effect of some specific event on people.

V. OBJECTIVES

- To carry out comparative study of different big data analytics techniques for unstructured big data.
- Design and form a better sentiment analysis technique for unstructured big data.
- Generalization of sentiment analysis technique.
- While considering multiple parameters, with accuracy, expected to be fast, precise and improved. It will help to design the strategies and reduce the business loss.
- The techniques will be generalized, useful not only for Indian Educational System, but for any area wherever voluminous data is required to be accessed within shortest time.

VI. RESEARCH METHODOLOGY

The researcher has planned to follow Design and Creation **r**esearch Strategy (figure no. 2). The strategy focuses on formation of new sentiment analysis technique for big data analytics.

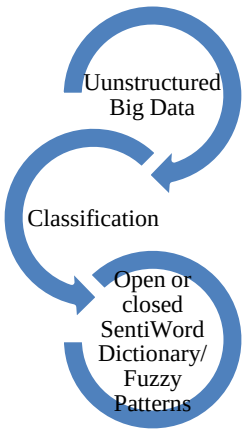


Figure No. 2. Collection and classification of Big Data

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

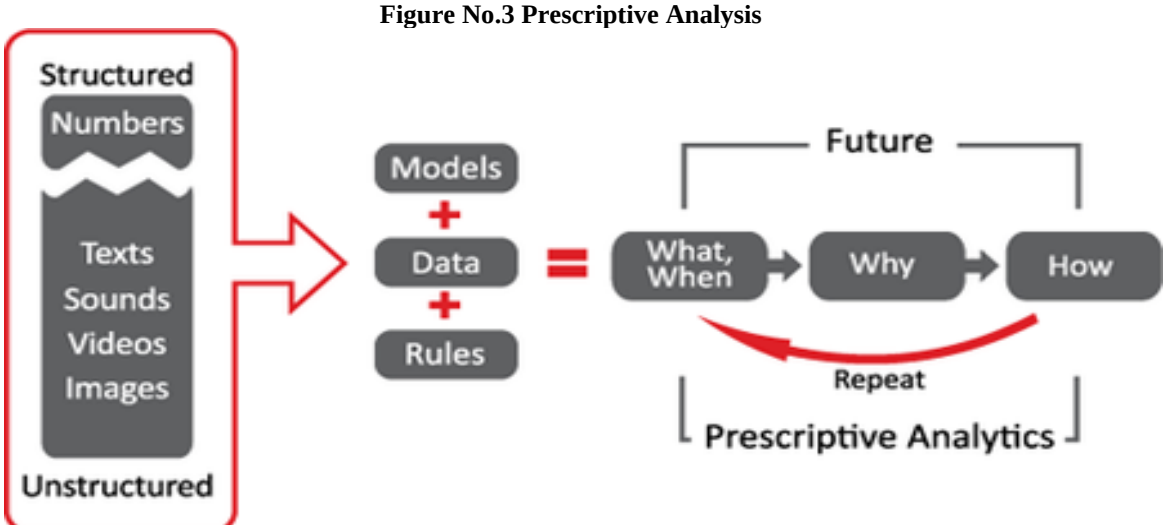
(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

The above Figure No. 2 shows the manner of pre-processing of unstructured big data and classification so as to apply phase wise algorithmic training, data specification and pattern recognition or to check fitness of membership for specified pattern ranges.

Unstructured data is data that does not follow a specified format for big data. Machine generated unstructured data includes Satellite images, scientific data such as seismic imagery, atmospheric data, and high energy physics, Photographs and video as security, surveillance, and traffic video, Radar or sonar data (vehicular, meteorological, and oceanographic seismic profiles) and human-generated unstructured data includes Text internal to your company all the text within documents, logs, survey results, and e-mails. Enterprise information actually represents a large percent of the text information in the world today, Social media data is generated from the social media platforms such as YouTube, Facebook, Twitter, LinkedIn, and Flickr, Mobile data (text messages and location information), website content: unstructured content, like YouTube, Flickr, or Instagram. Organisations stires such data and some organisations provides education, research and best practices to handle big data.However the technology didn't really support doing much with it except storing it or analysing it manually.

Figure No.3 Prescriptive Analysis



As shown the figure no. 3, Prescriptive analytics automatically synthesizes big data, multiple disciplines of mathematical sciences and computational sciences, and business rules, to make predictions and then suggests decision options to take advantage of the predictions. So as shown in Figure no. 3, collection of structured and unstructured big data includes challenges as capture, analysis, search, sharing, storage, transfer, visualization, and privacy violations etc. Researcher would like to apply technique to classify data with range partitions and instead of manual analytic efforts, formation of algorithmic technique to get synthesized data to be compared with fuzzy patterns.

Figure No. 3 shows that, it needs to simplify the large and complex data sets into smaller but logically related container sets, so all documents firstly will be divided into opinionated and non-opinionated documents, so as to focus only opinionated documents to synthesize furthermore. The researcher would like to apply intelligent models to this synthesis, by applying data mining or fuzzy based rules to get prediction of categories.

Then subjective and objective sentences of documents can be compared using automated rule based system to make predictions and to suggest decisions. Predicted results can be classified into three categories as positive, negative and neutral set of opinions.

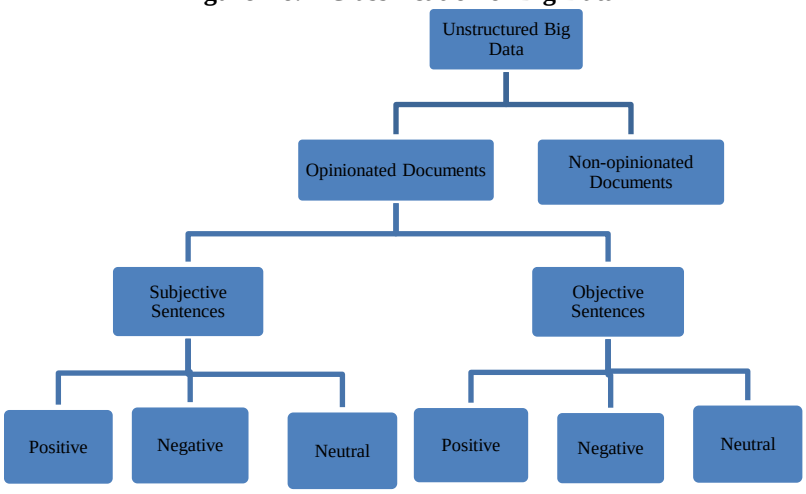
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

Figure No. 4 Classification of Big Data



By setting fuzzy weights to get identical features extraction and comparison for parameter passed as input. If sample data is collected, filtered and classified for datasets for analytics range partitions can help to form dataset in meaningful manner and then for every data set parameter membership for fuzzy specification can be evaluated to apply membership based rule to predict it's category of opinion (positive, negative, neutral) more accurately as shown in figure no.4.

By applying algorithm to datasets which are made based on range partitions for its identical features, one can identify and derive related opinion category.

Algorithm to give Input and to get Output predictions

Step 1: Perform Extract, Load and Transfer data tasks on data warehouse

Step 2: Develop Patterns and analysis

Step 3: Train pattern using Artificial Neural Network

Step 4: Apply Rule based fuzzy logic to patterns.

Step 5: Input: Extract parameters from structured and unstructured data by using data cleaning and data pre- processing.

Step 6: Process: Execute fuzzy model to apply fuzzification for identifying membership of given input parameter (i.e. check fitness approximation).

Step 7: Process: Apply inference rule according to comparison of identified member category.

Step 8: Output: based on training decide analytics result for matching input parameter from trained network.

VII. CONCLUSION

In this paper, by using smart intelligent strategy by applying ANN and fuzzy logic algorithm is proposed for sentiment analysis techniques. The proposed algorithm simplifies the challenges of unstructured and structured data pre-processing and Artificial Neural Network addresses to pattern learning and fuzzy logic helps to determine fitness approximation comparing for proper decision base for different ranges.

Proposed smart algorithm simplification and improvement in processing logic and speed apply to Big Data analytics and testing of algorithm for different business purposes may be future research domain. Significant smart sentiment analytics can help to predict habits, to take improved decisions, to recognise and design pattern for products and services, to design organisational business policies.

REFERENCES

1. Edd Dumbill, Big Data 2012 Edition O'Reilly, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472
2. Jalaj S. Modha, Prof. and Head Gayatri S. Pandi, Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and software Engineering, Volume 3, Issue 12, December 2013 pp no (91-97)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 12, December 2014

3. Meena Rambocas, João Gama "Marketing Research: The Role of Sentiment Analysis", FEP Working Papers, April 2013 ISSN: 0870-8541
4. Felipe Bravo-Marqueza, Marcelo Mendoza, Barbara Poblete "Meta-Level Sentiment Models for Big Social Data Analysis", Knowledge Based Systems May 2014
5. Demystifying Big Data: A Practical Guide to Transforming the Business of Government, TechAmerica Foundation's Federal Big Data Commission, 2012
6. George Gilbert, A guide to big data workload management challenges, May 2012, by Datastax.
7. Michael Kozlowski, "How big data helps the Education System, Jan. 2010.
8. http://en.wikipedia.org/wiki/Prescriptive_analytics
9. http://en.wikipedia.org/wiki/Big_data
10. http://en.wikipedia.org/wiki/Artificial_intelligence

BIOGRAPHY

Manisha Shinde-Pawar received the B.Sc. in Computer Science degree in 2003 and MBA degree in Information of Technology and Management in 2008. . She has also received Master of Computer Application (MCA). She has joined BVDU,IMRDA, Sangli, MS, India as Assistant Professor in Information Technology. Her research interests are distributed systems and mobile computing, big data etc.