e-ISSN: 2349-9745
p-ISSN: 2393-8161
Scientific Journal Impact Factor (SJIF): 1.711
International Journal of Modern Trends in Engineering and Research
www.ijmter.com

Text Analytics for Big Data

Sharvari Tamane
Department of Information Technology, MGMs Jawaharlal Nehru Engineering College, Aurangabad

Abstract— Most of the data used in various application areas like government, business and research is available in the form of text and therefore it is the requirement of these applications that it should derive high quality information by converting text into data for analysis purpose. The process of deriving high-quality information from the text is known as text analytics. Text analytics techniques represent knowledge, facts, business rules and relationships which are otherwise available in textual form incomprehensible for automatic processing. This paper mainly explores on how the different types of unstructured data are analyzed to get real meaning from data and which different text analytics tools are available for big data infrastructure. Routinely statistical and natural language processing techniques are used in text analytics to retrieve information from unstructured data. The idea behind this type of analytics is to determine who did what to whom, when, where, how and why. This information is then combined with structured information available in the data warehouse using various tools to gather further insight. At the end an overview of some of the players of this market is provided.

Keywords- Text Analytics, Big Data, Unstructured Data

I. INTRODUCTION

A. Unstructured Data

Eighty percent of data used in various documents, business information, customer correspondence, e-mails, social sites and/or any other sources which are important to particular organizations are unstructured in format. Unstructured data means the data which does not have proper predefined format. The volume and variety of this data are increasing in size day by day which makes it difficult to understand through general computer programs and to analyze it later. As many business decisions could be taken on the information extracted from unstructured data, users have approached to various software companies to retrieve the meaningful information from the unstructured data. These companies also help them to store and manage this data without knowing its format. Following are the examples of unstructured data [3, 4] with different characteristics: document files, spread sheets, digital data (audio/video), emails, graphics files etc.

Document files: This includes either static (which does not change frequently) or dynamic data. Static data includes: faxed data, scanned data, scientific data, resultant data, printed documents etc. Dynamic data includes: created data, office documents, reviewed data etc.

Spreadsheet: A spreadsheet stores data in terms of rows and columns. It is an interactive application used to organize and analyze data. Spreadsheets developed as computerized simulations of paper accounting worksheets. The data stored in the rows and columns may be either in structured, semi-structured or unstructured format. The formulae can be used to calculate cell values automatically. It also calculates and displays a value based on the contents of other cells.

Digital Data: This type of data has specialized technical requirement to store and manage the data for improving the performance. For example: Security/surveillance video, YouTube.

Emails: This includes data which is used to communicate between various users. This type of data

can also be treated as a business data.

Graphics files: This includes either machine or human generated data. Ex. Weather data, Satellite images, Google earth images, location information etc.

Web data (which is used to create, manage and maintain a web site), social media data (Facebook, Twitter etc.), data with associated intellectual property rights and Business Records (Documents, paper or e-business data which can be controlled over storage, retention, disposition and deletion to comply with legal, regulatory or industry requirements) are also known as unstructured data.

Parent Mr. X called to Class A. Enquired about class A.
Parent thinks feedback from previous batch should be taken.

Parent Mr. Y called to Class A. Enquired about class A.
It was ridiculous that the test series were not included in the teaching plan.

Potential called about Class A. Said that Class A fees was expensive.

Potential called about Class A. Said that having only two teachers is not sufficient.

Parent Mr. Z enquired about Class A. Said that teachers were not highly qualified.

Figure 1. Sample Class A Feedbacks

B. Text Analytics

Various methods are available to analyze unstructured data. These techniques [1] came out of technical areas such as Natural Language Processing (NLP), knowledge discovery, data mining, information retrieval and statistics. As mentioned in the abstract of this paper, the process of deriving high-quality information from the unstructured text is known as text analytics. Text analytics techniques represent knowledge, facts, business rules and relationships which are otherwise available in textual form incomprehensible for automatic processing. For example: The unformatted text from the classes may give you some details as to why this has happened? Figure 1. shows some of the "Class A" feedbacks. The underlined words provide information that why Class A is not known Class. An entity Class A appears throughout the feedbacks, indicating that the reports mentioned about the class. The terms "feedback from previous batch", "having only two teachers", "Class A" and "expensive" are evidences that, issues exist with feedback and the fees. Words like "ridiculous" and "not highly qualified" provide insight into the caller sentiment, which in this case is negative.
Different algorithms are available to process text analytics. The structured data extracted from the unstructured text is illustrated in Table 1. The difference between the "search" and "text analytics" is that search finds the words which users already know and text analytics is about discovering the information. But the later can augment search techniques.

C. Big Data

Big Data [5] is not a single technology but a combination of old and new technologies that helps companies gain actionable insight. Therefore, big data is the capability to manage a huge volume of disparate data at the right speed and within the right time frame to allow real-time analysis and reaction. Big data is typically broken down by three characteristics: Volume (i.e. how much data it has?), Velocity (i.e. how fast that data is processed?) and Variety (i.e. the various types of data: structured or unstructured?) as shown in figure 2. Even more important is the fourth V: veracity: (i.e.

how accurate is that data in predicting business value?). In a simple language big data is defined as it is so large and complex that it becomes difficult to process them using traditional data processing applications.

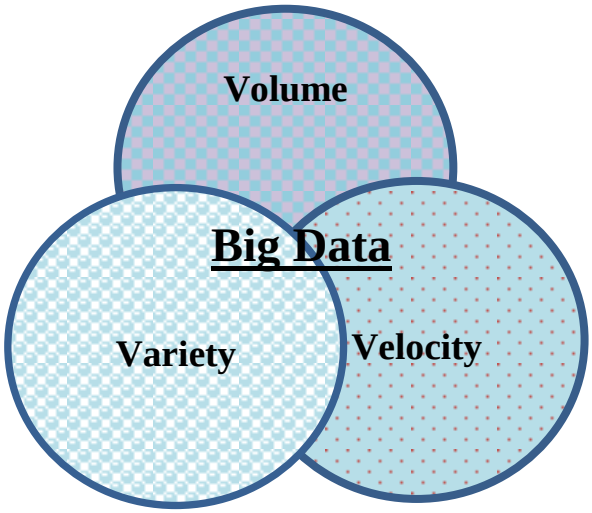

Figure 2. Big Data Characteristics

It is the requirement of the industry that it should be provided with useful insight and gain correct content, data should be processed with recent tools and produce important & meaningful information. An innovative business may want to be able to analyze massive amounts of data in real time to quickly assess the value of that customer and the potential to provide additional offers to that customer. It is necessary to identify the right amount and types of data that can be analyzed to impact business outcomes. Big data incorporates all data, including structured data and unstructured data from e-mail, social media, text streams etc. This kind of data management requires that companies leverage both their structured and unstructured data and may use different software tools to organize and manage these types of data.

II. TEXT ANALYTICS AND EXTRACTION TECHNIQUES

Generally, the text analytics methods [2] uses either statistical or Natural Language Processing (NLP) or a combination of statistical and NLP techniques to extract information from unstructured data. NLPs are used to obtain meaning from text. NLP commonly makes use of linguistic concepts such as grammatical structures and parts of speech. Many times these types of analytics are used to determine who did what to whom, when, where, how, and why.

Most common research tasks [2] used in NLP includes Co-reference resolution, Natural Language understanding, sentence breaking, speech recognition and many more.

In a co-reference resolution method a sentence or a paragraph may be used to search which phrases are used to refer the same objects. This may use language grammatical terms like nouns, pronouns etc. for reference purpose. It also identifies referring association for the phrase. For example: “Mr. XXX left the Prozone maul through the gate number 5”. “the gate number 5” is a referring phrase and the association to be identified is the fact that the gate being referred to is the gate number 5 of Prozone maul rather than any other structure which might also be referred to.

Natural language understanding method converts set of words into the words which are easy to manage for computer programs. This formation of semantic words can be done explicitly by constructing implicit assumptions such as subjective Yes/No can be constructed as an objective True/False.

Given a set of text, the sentence boundaries (dots or other punctuation marks) are also be used to

mark as an abbreviation. This method is known as sentence breaking or sentence boundary method.

In the speech recognition method a sound clip of a person determines the textual representation of the speech. Text to speech method works exactly opposite of this. Speech recognition method also uses speech segmentation and co-articulation processes to determine pauses between successive words and to remove blended letters from the speech respectively.

As mentioned in the reference [1] analysis on text through NLP can be performed at various levels. These levels are Lexical/morphological analysis, syntactic analysis, semantic analysis and discovered level analysis.

In Lexical/morphological analysis level characteristics of individual words are examined and then its meaning in the context of the text is provided. These words are examined by considering its prefixes, suffixes etc. along with its grammatical meaning (noun, verb, adjective etc.). These methods use dictionaries or publications usually in the form of a book, which provides meaning of a particular word. It also provides an application to discover various versions of that word. For example: demote, demotions and demoting are all versions of the same word.

Table 1. Structured Text from Unstructured Text

Variable	Entity	Remark	Sentiment
Parent Mr. X	Class A	Feedback from Previous batch	Neutral
Parent Mr. Y	Class A	Test series	Negative
XXXXX	Class A	Expensive	Neutral
XXXXX	Class A	Teachers	Negative
Parent Mr. Z	Class A	Teachers	Negative

In syntactic analysis level it uses syntactical or grammatical structure to separate the text and put each word into context. In this level a single word or a set of words can also be analyzed.

This step can be handled carefully to avoid incorrect sequences of words to make correct sentences. For example: The phone call records included the comment that “It was ridiculous that the test series were not included in the teaching plan”. Syntactic analysis would tag the noun phrases in addition to providing the part-of-speech tags.

In semantic analysis and discovered level analysis the possible meanings of a sentence and the meaning of text beyond the sentence level can be determined resp. Semantic analysis may include examining word order and sentence structure and disambiguating words by relating the syntax found in the phrases, sentences, and paragraphs.

To retrieve meaningful information from different sources companies sometimes requires developing simple/complex rules. For example: the first letter of a name of a person should be in capital letter, every course on the university website should contain alphabets and numbers, etc. Companies can generate these rules either manually or automatically or combination of both.

III. UNDERSTANDING THE EXTRACTED INFORMATION

The extraction of the text from various text analytics techniques automate the extracted information with tagging and markup which includes following types of information: terms (or keywords), entities (name (of a person/a company etc.), dates, time, Geographical locations, currencies, titles etc.), facts (relationships between two entities), events (same as facts except time dimension), concepts (Sets of words and phrases that indicates a particular idea or topic with which the user is concerned. This can be done manually or by using statistical, rule-based, or hybrid approaches to categorization.), and sentiments (to find emotions in the underlying text).

Once the unstructured data is made structured, this information is then combined with structured information available in the data warehouse using various tools to gather further insight. For example: in Table 2. text analytics results are merged with structured information. The contents of Table 2. are the same as Table 1., except a Segment column. This example shows that combining

structured data with the unstructured data makes some of the parents (customers) as gold customers so it would be worthwhile for the institute to make extra efforts to retain them. Of course, in reality, you will have a lot more data than this to work with.

Table 2. Combining Structured and unstructured data

Variable	Entity	Remark	Sentiment	Segment
Parent Mr. X	Class A	Feedback from Previous batch	Neutral	Gold
Parent Mr. Y	Class A	Test series	Negative	Silver
XXXXX	Class A	Expensive	Neutral	XXX
XXXXX	Class A	Teachers	Negative	XXX
Parent Mr. Z	Class A	Teachers	Negative	Bronze

If the data used in the above example is big data then the unstructured data being analyzed is either high volume or high velocity or both.

IV. TEXT ANALYTICS TOOLS FOR BIG DATA

This section provides an overview of some of the text analytics tools [1, 6].

Attensity: This is a text analytics and text mining tool which uses NLP to extract information from unstructured data. This is one of the original text analytics companies which started developing and selling products for more than 10 years. Attensity offers several engines (Auto-Classification, Entity Extraction, and Exhaustive Extraction) for text analytics. Exhaustive Extraction automatically extracts facts from parsed text (who did what to whom, when, where, under what conditions) and organizes this information. The company is focused on social and multichannel analytics and engagement by analyzing text for reporting from internal and external sources and then routing it to business users for engagement. It has developed a grid computing system that provides high performance capabilities for processing massive amounts of real-time text. Attensity uses a Hadoop framework (MapReduce, HDFS, and HBase) to store data. It also has a data-queuing system that creates an orchestration process that recognizes spikes in inbound.

Clarabridge: This is a text analytics and text mining tool which uses NLP, Machine learning, clustering and categorization techniques to extract information from unstructured data. Its goal is to help organizations to drive measurable business value by studying the customer of the whole. It also helps in taking important real time decisions depending upon customer feedback and staging the word to word for future processing into the Clarabridge system. It also offers its solution as a Software as a Service (SaaS).

IBM: IBM provides NLP solutions for analysis. This captures knowledge into dictionaries first and then semantic rules are applied to reuse it to extract customized information. The analyzed information can be analyzed again in the same way as the structured information would be analyzed. IBM Content Analytics (IBC) and Enterprise Search (ES) were once two separate products. The converged solution targets both enhanced enterprise search that uses text analytics, as well as standalone content analytics needs.

Open Text: This is a Canadian based company and famous for leadership in enterprise information management. This tool provides a platform to manage, capture, secure etc. unstructured data of enterprises.

SAS: it provides a broadly text analytic tool to extract information from unstructured/structured text. This tool uses statistical/NLP/advanced linguistic technologies to extract the text. The main

objective of this tool is to solve problems in minutes or days for which earlier it required days or week's resp. SAS analysis may be applied for thousands/millions of variables/documents. This text analytics are used in various areas like healthcare, digital content performance, high performance analytics etc. The solution runs on EMC Greenplum or Teradata appliances as well as on commodity hardware using Hadoop Distributed File System (HDFS).

***Ai-one*:** This tool provides APIs that enable developers to build machine learning applications.

V. CONCLUSION

Big Data is a collection of huge amount of data having different formats that it becomes very difficult to extract meaningful information using traditional data processing applications. Text analytics technology derives high quality information from unstructured text. Text Analytics uses statistical and NLP techniques to analyze the data. IBM, SAS, Attensity, Open text and clarabridge are some of the key vendors for text analytic tools.

REFERENCES

[1] S Judith Hurwitz, Alan Nugen, "Big Data for Dummies" John Wiley & Sons, Inc, 2013.
[2] Text Analytics, Natural Language Processing, Wikipedia, the free encyclopedia.
[3] Jose Carlos Gonzalez, Daedalus, Text Analytics 2014: April 12, 2014.
[4] Sumanda's Pages, My views on some current issues in Information space, Types of Unstructured Data, April 18, 2011.
[5] Sharvari Tamane, NoSQL Data Management for Big Data, Pre Science Congress: A National Level Conference on Science & Technology for Human Development, 30-31 December 2014, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India.
[6] Imanuel, Top 37 software for Text Analysis, Text Mining, Text Analytics, Text Analytics Software, TAGS, Random, Software.

