|  |  |  |  |
| --- | --- | --- | --- |
| **Method** | **Description** | **Example** | **Limitations** |
| Sentiment Analysis | Predicting the mood of the sentence like whether it is positive or negative [1]. | *“Product XYZ is good but expensive”*  The above statement states two aspects of the product XYZ where “product is good” shows the positive or favorable statement and “product XYZ is expensive” shows the negative or unfavorable statement [3]. | 1. Context – A decision cannot be made based on the words that are used in the context, as there will be two different meaning for the same word [5]. 2. Regional Variations – Language used in the context is a major limitation as different language words have different meaning [5]. |
| Optical Character Recognition in Natural Language Processing | Extracting useful information from the given image [1]. | Optical Character Recognition is used in Banking sectors to process the checks, just a scan of the check will process the transaction without any human involvement and it is also used in other industries like finance, education and legal industries to digitize the records or documents [4].  OCR it is also used for the automatic number plate reading and also used as an aid for blind people [2].  OCR is especially used to extract information from the unstructured data as the data can be in any format like image, text, graphics etc.  In OCR the text has been scanned, preprocessed, segmented and data has been extracted. | 1. The main limitation in extracting information is because the data is often mixed with text and graphics [2]. 2. Variations in style and shape of the data [2]. 3. Variations because of subscripts and superscripts in the data [2]. |
| Data Mining in Information Retrieval | Process of analyzing and extracting useful information from different perspectives by using various data mining concepts such as clustering, classification etc. [6]. | Data mining is applied in Healthcare Industry for the evaluation of treatment effectiveness, by comparing the causes, symptoms, and course of treatments [7].  Data mining is basically retrieving information from the different types of data. The data will be in any format, in healthcare industry the data can be in any format such as medical prescription, X-ray or scan reports. All these are analyzed and valuable information is extracted. | 1. Issues may arise with the missing, corrupted, inconsistent data as the information recorded will be in different format from different sources [7]. |
| Text Analytics | Text analytics helps in retrieving the valuable text information from the unstructured and semi-structured data. Text analytics is also referred as text mining [9]. | Text analytics or text mining is used in many fields like Publishing and media, Banks, Telecommunications etc. [11]. | 1. Lot of software programming is needed to extract textual information from different sources [9]. 2. Managing the unstructured data from various sources is complicated [9]. |
| Big Data Analytics | Big data refers to large or complex data that are difficult to manage with the traditional software, hardware and data management tools [10]. | Big data analytics is used in many industries. For instance, in the health care industry it is used to detect the disease at the earlier stage where they can be treated more easily and effectively [10]. | 1. Data extraction and cleaning is one of the limitation in big data analytics. 2. Data Integration is another limitation as the data will be coming from different databases or web portals. |
| Auto completion in NLP | Predicting the word which follow the sequence of words [1]. | For Example, the word “Blue” is more likely to follow the sequence “The sky is” [1].  This auto completion is used to convert the unstructured data to structured data.  For instance, creating an abstract automatically by reading a document or a word following a sequence. |  |
| Part-of-speech tagging | Tagging the part of speech for each word in a sentence such as noun, verb, adverb, etc. [1]. | *For example, “a cotton shirt” this words are tagged as Nouns [16].*  Part of speech tagging is used in speech recognition and information retrieval [15].  In Speech recognition POS helps in recognize the pronunciation. It also helps in preprocessing. | 1. *The main challenge in the POS is removing ambiguities, for instance “Flies like a flower” here the POS for the words in a sentence are*   *Flies: noun or verb?*  *like: preposition, adverb, conjunction, noun, or verb?*  *a: article, noun, or preposition?*  *flower: noun or verb?*[[1]](#footnote-1) |
| Web Crawling in Information Retrieval | Automatic Script that can download the contents from the World Wide Web automatically [6]. | Web crawling in Information Retrieval used to retrieve information from various sources and this web crawler is one of the important component in search engines which will help us to extract valuable information from the web pages [6]. | 1. Web crawling is very difficult because of its large volume and its rate of change [14]. 2. Large volume states that it can download only certain limit of web pages at a time. 3. Rate of change implies that the web page either might be added or deleted as there will be lot of changes in the web pages on daily basis. |
| Text Summarization in NLP | Summarizing the whole content of the text document in a shorter version without changing the overall meaning of the document [11]. | It is helpful in summarizing single and multi-document by extractive or abstractive text summarization method [13].  The documents from various sources will have lengthy and unstructured text, this text summarization helps in converting and extracting useful information or structured text using two methods like abstractive and extractive. | 1. Summarizing the whole content of the document from various sources in a right way (language, format, etc.) to a specific user [11]. |
| Stemming in Natural Language Processing | Stemming is a process of deriving a word to their root [17].  *For instance, the word decided(Adjective), decision(Noun) is basically derived from the word “decide” so here the word decide is a root word or stem word [17].* | Stemming is an important feature in today’s search systems. The main idea of this stemming process is reducing the word to their root while searching [18].  Stem word usually represents broader concept than the original term which will be helpful in retrieving large number of documents [18]. | 1. Over stemming is where the two different words are stemmed to the same root [18]. 2. Under stemming is where the two different words should be stemmed to the same root but not [18]. |
| Indexing in Information Retrieval | Indexing in Information retrieval is known as the process of extracting information without analyzing the content [6]. | *The main application of the indexing is to recognize the contents of the document.[[2]](#footnote-2)* |  |
| Topic Segmentation in Natural Language Processing | Separating a given text into the specific topic [1]. | *It helps in retrieving information from the set of words appear in one or more paragraphs which is likely to contain the definition of the word.[[3]](#footnote-3)* |  |
| Boolean Model in Traditional Information Retrieval | Boolean model in information retrieval is based on the Boolean Algebra. It represents the model by set of index terms which are viewed as Boolean variable and valued as True if it is present in document [19]. | *For instance, consider we have three documents and they are [20]*  *Doc1: Information Retrieval has 2 models and Information.*  *Doc2: Boolean is a basic Information Retrieval classic model.*  *Doc3: Information is a data that processed, Information.*  *If the Query for above documents be*  *(Data ^ Information) V (~ Retrieval)*  *Result:*  *Data: Doc3*  *Information: Doc1, Doc2, Doc 3*  *Retrieval: Doc1, Doc2*  [20] | 1. Retrieval performance is very poor [19]. |
| Vector Space Model in Traditional Information retrieval | It is an Algebraic model where the documents are represented as vectors. Each vector represents a word or keyword associated with the document [20]. | *Example in [20].* | 1. It is difficult for the vector space model to conclude the words resembling the same concept [1]. |
| Probabilistic Model in Traditional Information Retrieval | Probabilistic model attempts to estimate the probability of the user finding a document. The ranking in this model is based on the documents retrieved to the given query [20]. | *The information is extracted based on the query which has been passed by the user. Based on the query the best results are listed out first [20].* | 1. Probabilistic models are very hard to build and program. |
| Audio Analytics in Big data analytics | Audio analytics analyze and extract information from unstructured audio data. This is also referred to as speech analytics as this technique is used in speech recognition as well [21]. | *Audio analytics are used in customer call centers which will help to improve customer experience, enhance sales turnover rates, etc. [21]* | It is difficult to have 100% accuracy in the identification of audio stream; it may not be able to handle accented words. Sometimes understanding the contextual meaning of the words would still be a challenge.[[4]](#footnote-4) |
| Video Analytics in Big data analytics | Video analytics analyze and extract meaningful information from video stream [21]. | *Video analytics is mostly used in security and surveillance systems. Video analytics can easily detect breaching, loitering, theft etc.* [21]. | There are three approaches to analyze video:   * Server-based (the videos which are captured through cameras are routed back to a centralized server to perform analysis) * Edge based (the video analysis is performed locally or on raw data captured by the cameras) * Agent Vi’s distributed architecture (in this the analysis work is distributed between the edge device and the server).   The accuracy of server-based analysis is less as the data is generated in compressed form due to limited bandwidth, but it facilitates easier maintenance while in edge based the entire content is available for analysis give better results but it is costly to maintain and have lower processing power as compared to server-based analysis [22]. |
| Social media analytics | Analyzing structured and unstructured data from the social media channels is referred to as social media analytics.  Social media are such as social networks like Facebook, Twitter and blogs like WordPress and other platforms like Instagram YouTube etc. [21]. | *The main application of social media analytics is the marketing field where the companies can analyze the reaction of the people of their products [21, 22].* | 1. Massive amount of data requires lot of storage space and processing power [23]. 2. Worldwide online accessibility provides more data in many languages [23]. |
| Predictive analytics | Predictive analytics is a method where the useful information is extracted from the existing data, basically it doesn’t tell what will happen in the future instead it forecasts what might happen in the future [23]. | *Predictive analysis is used in Customer relationship management fields like marketing, sales, customer services etc. to analyze the product in demand and predict the customer buying habits [23].*  *It is also used in clinical decision support to predict whether the patients are at risk of developing certain conditions like asthma, diabetes etc. [23].* | 1. Predictive analysis technique is based on statistical methods but more statistical methods must be developed for big data as they are massive [22].   *Statistical methods are methods of collecting, summarizing, analyzing, and interpreting variable numerical data.*[[5]](#footnote-5) |

**REFERENCES**

1. [Gabriele Bavota](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Gabriele%20Bavota.QT.&newsearch=true), Mining Unstructured Data in Software Repositories: Current and Future Trends, [Software Analysis, Evolution, and Reengineering (SANER), 2016 IEEE 23rd International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7476313) 14-18 March 2016

# Line Eikvil, Optical Character recognition, December 1993

1. Tetsuya Nasukawa, Jeonghee Yi, Sentiment Analysis: Capturing Favorability using Natural Language Processing
2. Ravina Mithe, Supriya Indalkar, Nilam Divekar, Optical Character Recognition, International Journal of Recent technology and Engineering, March 2013

# [V Nagarjana Devi Duvvuri](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V%20Nagarjana%20Devi%20Duvvuri.QT.&newsearch=true), [V.Rajanikanth Thatiparthi](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.Rajanikanth%20Thatiparthi.QT.&newsearch=true), [Rajashekar Pantangi](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Rajashekar%20Pantangi.QT.&newsearch=true), [Akhil Gangavarapu](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Akhil%20Gangavarapu.QT.&newsearch=true), Sentiment Analysis Using Harn Algorithm, [IT Convergence and Security (ICITCS), 2016 6th International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7740253) 26-26 Sept. 2016

# [Chandni Saini](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Chandni%20Saini.QT.&newsearch=true), [Vinay Arora](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Vinay%20Arora.QT.&newsearch=true), Information retrieval in web crawling: A survey,  [Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7592392) 21-24 Sept 2016

# Hian Chye Koh, Gerald Tan, Data Mining Applications in Healthcare

# Namrata H.S, B.S. Satpute, Pramod Patil, Web Forum Crawling Techniques, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014

# Kalli Srinivasa Nageswara Prasad, Prof S. Ramakrishna, Text Analytics to Data Warehousing.

# Wullianallur Raghupathi and Viju Raghupathi, Big data analytics in healthcare: promise and potential, 07 Feb 2014

# Vishal Gupta, Gurpreet S.Lehal, A survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, August 2009

# Saumya Salian, Challenges with Big Data Analytics, International Journal of Science and Research

# Udo Hahn, Inderjeet Mani, The challenges of Automatic Summarization, November 2000

# Carlos Castillo, Effective Web Crawling, November 2004

# Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging

# Beatrice Santorini, Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision), July 1990

# [Rahul S. Dudhabaware](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Rahul%20S.%20Dudhabaware.QT.&newsearch=true), [Mangala S. Madankar](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Mangala%20S.%20Madankar.QT.&newsearch=true), Review on Natural Language Processing Tasks for Text Documents,  [Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7203316), 18-20 Dec. 2014

# Anjali Ganesh Jivani, A comparative study of Stemming Algorithms. November 2011

# [V.N. Gudivada](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.N.%20Gudivada.QT.&newsearch=true), [W.I. Grosky](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.W.I.%20Grosky.QT.&newsearch=true), [V.V. Raghavan](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.V.%20Raghavan.QT.&newsearch=true), Information retrieval on the World Wide Web,  [IEEE Internet Computing](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/RecentIssue.jsp?punumber=4236), 06 August 2002

# [Arash Habibi Lashkari](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Arash%20Habibi%20Lashkari.QT.&newsearch=true), [Fereshteh Mahdavi](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Fereshteh%20Mahdavi.QT.&newsearch=true), [Vahid Ghomi](http://ieeexplore.ieee.org/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Vahid%20Ghomi.QT.&newsearch=true), A Boolean Model in Information Retrieval for Search Engines, [Information Management and Engineering, 2009. ICIME '09. International Conference on](http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=5076977) 3-5 April 2009

# Amir Gandomi, Murtaza Haider, Beyond the hype: Big data concepts, methods, and analytics, International Journal of Information Management

# Poonam Vashisht, Vishal Gupta, Big data analy-tics techniques: A survey, Green Computing and Internet of Things International Conference, 8-10 Oct 2015

# Jai Prakash Verma, Smita Agrawal, Bankim Patel and Atul Patel, Big data analytics: Challenges and applications for Text, Audio, Video and social media data, International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), February 2016

1. https://www.cs.umd.edu/~nau/cmsc421/part-of-speech-tagging.pdf [↑](#footnote-ref-1)
2. <http://www.iro.umontreal.ca/~nie/IFT6255/Indexing.pdf> [↑](#footnote-ref-2)
3. https://dias.users.greyc.fr/nlp/NLP-Chapter6-Topic-Segmentation.pdf [↑](#footnote-ref-3)
4. https://books.google.com/books?id=P5BJDAAAQBAJ&lpg=PA60&ots=KBG2jw7BEp&dq=limitations%20of%20audio%20analytics&pg=PA60#v=onepage&q&f=true [↑](#footnote-ref-4)
5. http://www.encyclopedia.com/computing/dictionaries-thesauruses-pictures-and-press-releases/statistical-methods [↑](#footnote-ref-5)