Text Analytics: the Convergence of Big Data and Artificial Intelligence

Swapnil S. Chopade[1] Aishwarya C. Chowdhari[2] Siddhi S. Chorghe[3] Ashwini D. Padekar[4]
[1,2,3]Student [4]Assistance Professor
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]MGM's College of Engineering and Technology, Kamothe, Navi Mumbai, India

Abstract— The analysis of the text content in emails, tweets, SMS, blogs, forums and other forms of textual communication is called text analytics. Text analytics is applicable to most industries: it can help analyse millions of emails, using text analytics we can analyse customer's comments and questions in forums, we can perform sentiment analysis using text analytics by measuring positive or negative perceptions of a company, brand, or product. Text Analytics has also been called text mining, and is a subcategory of the Natural Language Processing (NLP) field, which is one of the founding branches of Artificial Intelligence, back in the 1950s, when an interest in understanding text originally developed. Currently Text Analytics is often considered as the next step in Big Data analysis. Text Analytics has a number of subdivisions: Information Extraction, Semantic Web annotated domain's representation, Named Entity Recognition, and many more. Several techniques are currently used and some of them have gained a lot of attention, such as Machine Learning, to show a semi supervised enhancement of systems, but they also present a number of limitations which make them not always the only or the best choice. We conclude with current and near future applications of Text Analytics.

Key words: Big Data Analysis, Information Extraction, Text Analytics

I. INTRODUCTION

Natural Language Processing (NLP) is the practical field of Computational Linguistics, there are some authors who uses this term almost interchangeably. NLP has been considered a sub discipline of Artificial Intelligence, and more recently it is core of Cognitive Computing, since most cognitive processes are either understood or generated as natural language utterances. NLP is a very broad topic, and includes a huge amount of Subdivisions such as: Natural Language Understanding, Knowledge Base building, Natural Language Generation, Dialogue Management Systems and Intelligent Tutor Systems in academic learning systems, Speech Processing, Data Mining – Text Mining – Text Analytics, and so on. We will focus here in this specific article in Text Analytics (TA).*Text Analytics* is nothing but Natural Language Understanding (NLU), it is the most recent name given to Natural Language Understanding, Data and Text Mining. In the last few years a new name has gained popularity which is Big Data, Big data refer mainly to unstructured text (or other information sources), Big data is more often used in the commercial sectors rather than the academic sectors, these may be because unstructured free text accounts for 80% in a business context, including tweets, wikis, blogs, and surveys [1]. Text Analytics has become an important research area. Text Analytics is the discovery of new, previously unknown information, by automatically extracting information from different written resources.

II. TEXT ANALYTICS: CONCEPT & TECHNIQUE

Text Analytics is an extension of data mining that tries to find textual patterns from large non-structured sources, as opposed to data stored in relational databases. Text Analytics is also known by other names such as Text Data Mining or Knowledge-Discovery in Text (KDT), Text analytics refers generally to the process of extracting non-trivial information and knowledge from unstructured text. Text Analytics is similar to data mining, the only difference is that data mining tools are designed to handle structured data from databases which are either stored as such or as a result from pre-processing unstructured data. Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents, HTML files, blogs, newspaper articles, academic papers, etc. Text Analytics is gaining prominence in many industries, from marketing to finance, because the process of extracting and analysing large quantities of text can help decision-makers to understand market dynamics, predict outcomes and trends, detect fraud and manage risk. The multidisciplinary nature of Text Analytics is key to understand the complex integration of different expertise: computer engineers, linguists, experts in Law, Biomedicine etc.

A. Information Extraction

Information Extraction (IE) is a technique that extract meaningful information from large amount of text. Domain experts specify the attributes and relation according to the domain. IE systems are used to extract specific attributes and entities from the document and establish their relationship. The most popular form of IE is named entity recognition (NER). NER seeks to locate and classify atomic elements in text into predefined categories (usually matching pre-established ontologies). NER techniques extract features such as the names of persons, organizations, locations, temporal or spatial expressions, quantities, monetary values, stock values, percentages, gene or protein names, etc. Precision and recall process is used to check and evaluate the relevance of results on the extracted data. The recent activities in multimedia document processing like automatic annotation and mining information out of images/audio/video could be seen as information extraction and the best practical and live example of IE is Google Search Engine.

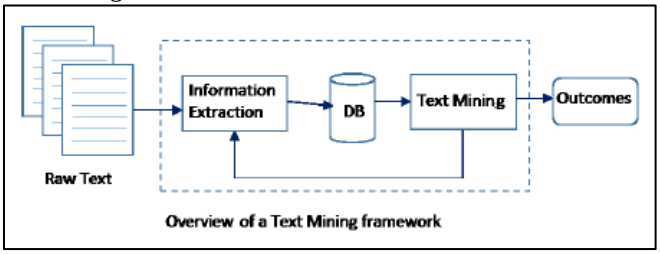

Fig. 1: Overview of a Text Mining Framework

B. Topic Tracking & Detection

Keywords are a set of significant words in an article that gives a high-level description of its contents to readers. Identifying keywords from a large amount of online news data is very useful in that it can produce a short summary of news articles. As online text documents rapidly increase in size with the growth of WWW, keyword extraction has become the basis of several text mining applications such as search engines, text categorization, summarization, and topic detection.

C. Summarization

Text summarization is a process of collecting and producing concise representation of original text documents. Pre-processing and processing operations are performed on the raw text for summarization.

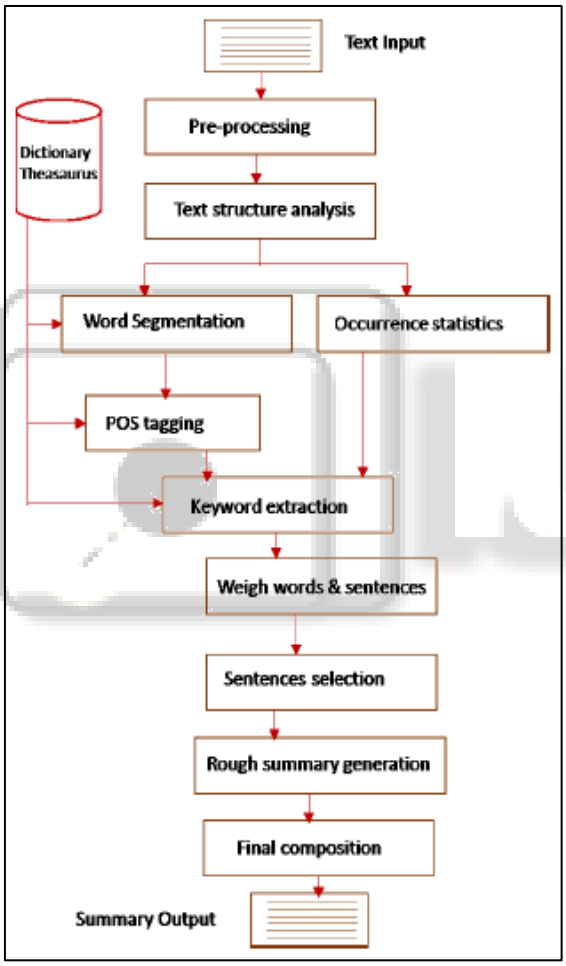

Fig. 2: Text Summarization

Tokenization, stop word removal, and stemming methods are applied for pre-processing. Lexicon lists are generated at processing stage of text summarizationText summarization has a long and fruitful tradition in the field of Text Analytics. In a sense text summarization falls also under the category of Natural Language Generation. It helps in figuring out whether or not a lengthy document meets the user's needs and is worth reading for further information. Text summarization techniques can be applied on multiple documents at the same time.

D. Clustering

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered without the use of predefined topics. In other words, while categorization implies supervised (machine) learning in the sense that previous knowledge is used to assign a given document to a given category, clustering is unsupervised learning: there are no previously defined topics or categories.

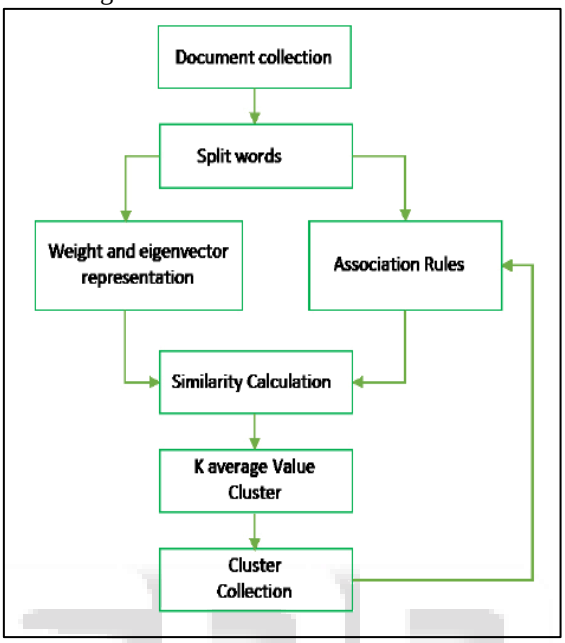

Fig. 3: Document Clustering

Using clustering, documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results (multiple indexing references). Clustering is an unsupervised process to classify the text documents in groups by applying different clustering algorithms. In a cluster, similar terms or patterns are grouped extracted from various documents. Clustering is performed in top-down and bottom up manner. In NLP, many types of mining tools and techniques are used for the determination on unstructured text. Various methods of clustering are distribution, density, centroid, hierarchical and k-mean. The use of K-Means algorithm allow us to implement semi supervised learning clusters using an algorithm so as to help identify approximate the text to search using predefined patterns and the implementation of a cluster algorithm [6] for consultations within the database manager MySQL in order to obtain scientific research papers.

E. Categorization or Classification

Categorization involves identifying the main themes of a document by placing the document into a predefined set of topics. Categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on relationships identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic [10]. Another method is to represent topics as thematic graphs,

and using a degree of similarity (or distance from the "reference" graph) to classify documents under a given category [11].

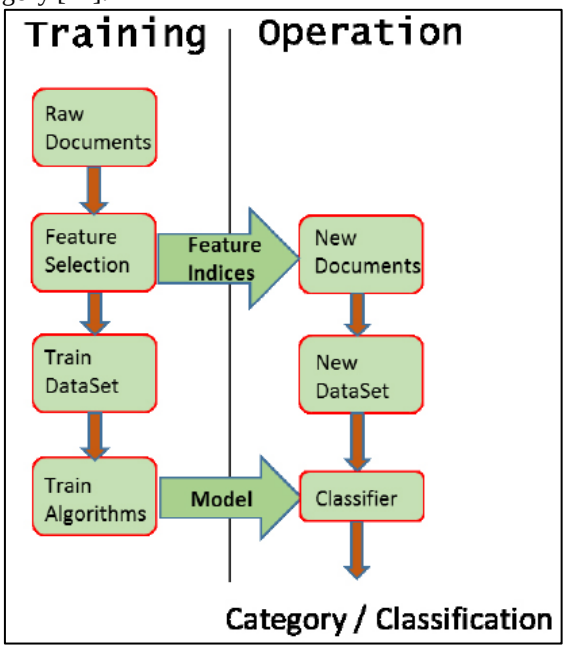

Fig. 4: Text Classification

III. Some Use Cases

A. Digital Libraries

Numerous text mining techniques and tools are in use to ascertain the patterns and trends from journals and proceedings from immense amount of repositories. These sources of information help in the field of research and development. Libraries are a great source of information for the researchers and digital libraries are endeavouring to the significance of their collection. It provides a novel method of organizing information in such a way that make it possible to available trillions of documents online. It provides a novel way to organize information and make it possible to access millions of documents online. Green-stone international digital library that support multiple languages and multilingual interfaces provide a springy method for extracting documents that handle multiple formats, i.e., Microsoft word, pdf, postscript, HTML, scripting languages and e-mail messages.

B. Life Science

Life science and health care industries are generating large amount of textual and numerical data regarding patient's record, diseases, medicines, symptoms and treatments of diseases and many more. It is a big challenge to filter out an appropriate and relevant text to take a decision from a large biological repository [25]. The medical records contain varying in nature, complex, lengthy and technical vocabulary are used that make the knowledge discovery process very difficult. Machine-based conclusions could help the public to handle the mass of information and medical experts to give expert their feedback. An instant classification of amateur requests to medical expert network forums is a heavy task because these requests can be long and unstructured as an end of mixing, for example, personal experiences with laboratory data. It is a big challenge to find out a correct and important text to take a right decision from a huge biological repository.

C. IBM's Watson

IBM has a long history in AI research, and they consider themselves as one of the founding fathers of AI in the early days in the 1950s. Along the years IBM created a checkers player ; a robotic system trained to assemble some parts of an IBM typewriter ; Deep Blue, the specialized chess-playing server that beat then World Chess Champion, Garry Kasparov ; TD-Gammon, a backgammon playing program using reinforcement learning (RL) with multiple applications ; and pioneering work in Neural Network Learning, inspired in biological information processing . More recently work focused on advanced Q&A (question and answer) and Cognitive Computing, of which the Synapse project and Watson are the more well-known examples.

IV. Examples of TA Applications

We will briefly review two prominent areas of application of Text Analytics, with a large commercial impact: (1) Medical Analytics – classification of articles of medical content, and (2) Legal Analytics– Information extraction from legal texts.

A. Medical Analytics – Classification of articles or medical content

Biomedical text mining or BioNLP presents some unique data types. Their typical texts are abstracts of scientific papers, as well as medical reports. The main task is to classify papers by many different categories, in order to feed a database (like MEDLINE). Other applications include indexing documents by concepts, usually based or related to ontologies or performing "translational research," that is, using basic biological research to inform clinical practice (for instance, automatically extraction of drug-drug interactions, or gene associations with diseases, or mutations in proteins). There are three approaches for extracting relations between entities:

- Linguistic-based approaches: the idea is to employ parsers to grasp syntactic structures and map them into semantic representations.
- Pattern-based approaches: these methods make use of a set of patterns for potential relationships, defined by domain experts.
- Machine Learning-based approaches: from annotated texts by human experts, these techniques extract relations in new collections of similar texts.

B. Legal Analytics – Information extraction from legal texts

One area getting a lot of attention about the practicalities of Text Analytics is that concerning the information extraction from texts with legal content. More specifically, litigation data is full of references to judges, lawyers, parties (companies, public organizations, and so on), and patents, gathered from several millions of pages containing all kinds of Intellectual Property (IP) litigation information. This has given rise to the term Legal Analytics, since analytics helps in discovering patterns with meaning hidden in the

repositories of data. What it means to lawyers is the combination of insights coming from bottom-up data with top-down authority and experience found in statutes, regulations and court sentences. While a search can be made for the string *plaintiff*, there are no searches for a string that represents an individual who bears the role of plaintiff. To make language on the Web more meaningful and structured, additional content must be added to the source material, which is where the Semantic Web (semantic roles' tagging) and Natural Language Processing perform their contribution. We start with an input, the corpus of texts, and then an output, texts annotated with XML tags, JSON tags or other mechanisms. However, getting from a corpus of textual information to annotated output is a demanding task, generically referred to as the knowledge acquisition bottleneck [43].

V. Future work

The technologies around text analytics are currently being applied in several industries, for instance, sentiment and opinion analysis in media, finance, healthcare, marketing branding or consumer markets. Insights are extracted not only from the traditional enterprise data sources, but also from online and social media, since more and more the general public has turned out to be the largest generator of text content (just imagine online messaging systems like WhatsApp or Telegram). The current state of text analytics is very healthy, but there is room for growth in areas such as customer experience, or social listening. This bears good promises for both scientific experimentation and technical innovation alike: Multi-lingual analytics is facilitated by machine learning (ML) and advances in machine translation; customer experience, market research, and consumer insights, and digital analytics and media measurement are enhanced through text analytics; besides the future of deep learning in NLP, long-established language engineering approaches taxonomies, parsers, lexical and semantic networks, and syntactic-rule systems will continue as bedrocks in the area; emotion analytics, affective states compounded of speech and text as well as images and facial-expression analysis; new forms of supratextual communications like emoji's need their own approach to extract semantics and arrive at meaningful analytics; semantic search and knowledge graphs, speech analytics and simultaneous machine translation; and machine-written content, or the capability to compose articles (and email, text messages, summaries, and translations) from text, data, rules, and context, as captured previously in the analytics phase.

VI. Conclusion

Text Analytics, with its long and prestigious history, is an area in constant evolution. It sits at the centre of Big Data's Variety vector, that of unstructured information, especially with social communications, where content is generated by millions of users, content not only consisting of images but most of the times textual comments or full blown articles. Information expressed by means of texts involves lots of knowledge about the world and about the entities in this world as well as the interactions among them. That knowledge about the world has already been put to use in order to create the cognitive applications, like IBM's Watson and IPsoft's Amelia that will interact with human beings expanding their capabilities and helping them perform better. With increased communication, Text Analytics will be expanded and it will be needed to sort out the noise and the irrelevant from the really important information. The future looks more than promising.

References

[1] P. Cowling, S. Remde, P. Hartley, W. Stewart, J. Stock-Brooks, T. Woolley, " C-Link Concept Linkage in Knowledge Repositories", Vol.01, September 2010.
[2] Q. Lu, J. G. Conrad, K. Al-Kofahi, W. Keenan," Legal document clustering with built-in topic segmentation", Vol.04, January 2011.
[3] H. Cordobés, A. Fernández Anta, L.F. Chiroque, F. Pérez García, T. Redondo, A. Santos, "Graph-based Techniques for Topic Classification of Tweets in Spanish", Vol.03, January 2008.
[4] T. Theodosiou, N. Darzentas, L. Angelis, C.A. Ouzonis, "PuReDMCL: a graph-based PubMed document clustering methodology". Vol.01, February 2008.
[5] G. Wen, G. Chen, and L. Jiang, "Performing Text Categorization on Manifold", Vol.04, March 2006.
[6] W. Xiaowei, J. Longbin, M. Jialin and Jiangyan, "Use of NER Information for Improved Topic Tracking", Vol.05, March 2008.
[7] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", Vol.02, June 2005.
[8] S. Lee and H. Kim, "News Keyword Extraction for Topic Tracking", Vol.01, October 2008.
[9] C. Friedman, T. Rindflesch, and M. Corn, "Natural language processing:State of the art and prospect for significant progress", Vol.02, February 2013.
[10] T. Redondo, "The Digital Economy: Social Interaction Technologies– an Overview", Vol.03-2, September 2015.
[11] Karl Flinders, "Amelia, the IPsoft robot", Vol. 2, September 2015.
[12] R. Linsker, "Perceptual Neural Organization: Some Approaches Based on Network Models and Information Theory", Vol.13, April 1990.
[13] G. Tesauro, "Temporal difference learning and TDGammon Gammon",Vol. 38, June 1995.

