International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Index Copernicus Value (2016): 79.57 | Impact Factor (2017): 7.296

# An Overview of Text Analysis Technologies and Tools

**Akshita Lakkad[1], Sneh Gajiwala[2], Dr. (Mrs.) Neepa Shah[3]**

[1]Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India

[2]Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India

[3]Head of Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai 400056, India

**Abstract:** ***Text analytics is the process of reading unstructured textual data and converting it into meaningful data for analysis, to obtain a measurable quantity that provides some valuable information. Text analysis is being increasingly adopted by business organizations. It helps business organizations to interpret unstructured data such as customer feedback and identify patterns and predict trends. Many software solutions for text analysis provide tools, servers, analytical algorithms based applications, data mining and extraction based tools for converting text data into meaningful data for analysis. In this paper, we have discussed the basics of text analysis, various techniques of text mining, and the most popular tools for text analysis. Examples of tools include natural language processing, high-end APIs that can be easily integrated with other software, full-text and keyword search.***

**Keywords:** text analysis tools, semantic analysis, word embedding, techniques, topic modeling, natural language processing

## 1. Introduction

In today's world, valuable information can be found anywhere and in any form. User reviews, newspaper articles, blogs etc. can all have information that can be relevant to you. Majority of this data is available in the form of text. Sometimes, when the need for precise information arises, it becomes difficult to parse through all the existing data. It becomes too much to read. Also, most of the times the information is buried and manual analysis of this abundant data is practically time-consuming. All the textual data in these documents is unstructured and fuzzy. This is where the text mining tools come into picture. These tools are used to extract relevant nuances of information, interpret the data to gain insights and patterns from the same. Thus, providing a structure to the formerly unstructured data[1].

Text Mining is a broad term comprising information retrieval, text analysis, text extraction, data mining, classification, categorisation and machine learning[2].

The two stages in text mining process include acquiring relevant information from unstructured text. Once we have the relevant information, the next stage is categorising this information in different ways. The extracted data can either be classified based on specific classes or arranged in the form of clusters. Text categorisation or text classification is the smart classification of text into predefined categories. Intent, emotions and sentiment analysis of the textual data are the most common tasks of for text classification. Text classifiers can operate on a variety of datasets and they can be trained on supervised or unsupervised data based on the data that is being classified.

Text in the same cluster tends to be more similar in nature as compared to the text in other clusters. These clusters can be based on the interpretations we are trying to gather from the data. For instance, if we are trying to form clusters for online reviews, there may be separate clusters for positive, negative and neutral reviews. These clusters can be used to interpret how the product is being received by the audience. Lastly, machine learning algorithms can be applied to extract patterns and insightful information from the data.[1]
The combination of these activities together forms the core of text mining. Text analysis is an emerging technology that is being adopted by various business organizations because of its increasing applications in developing a business strategy. There are various text analysis tools that are being developed to assist business organizations [3]. The popularity of such tools is growing multifold. The steps in text mining and analysis are depicted in the figure1.

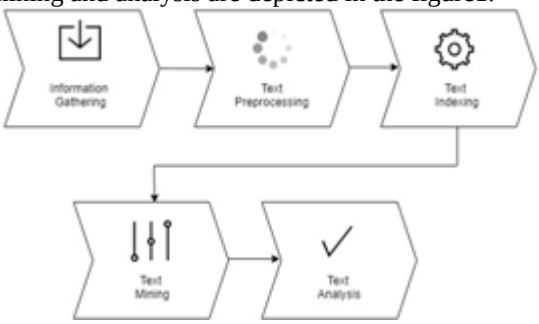


**Figure 1**: Steps in Text Mining

In this paper, we analyze the highly efficient and accepted techniques used for text analysis and compare the tools that use these techniques. We have discussed how each of these tools work and where their most appropriate uses might lie.

In Section 2 we discuss about the popular techniques used by the tools. Section 3 describes the tools based on the features and the services they provide. Section 4 and 5 describes the features and categories of these tools. Section 6 consists of a comparison of the tools. Lastly, section7 provides a conclusion.

## 2. Approaches to text analysis

Text mining tools are generally based on specific stand-alone or a combination of multiple techniques depending on the purpose these tools serve. These tools follow algorithms that form the basis of the entire analysis and mining procedure.

The most popular techniques and their percentages are shown in figure 2 below[1].

### 2.1 Natural Language Processing - Word Embedding:

A lot of Deep Learning models are based on numeric data. Processing of textual data in these cases is a prolonged process. Word embedding is used to represent text in numeric format. Thus, words are mapped to a vector. Commonly used word embedding algorithms are TF-IDF, Skip-Gram model, Common Bag of Words etc.

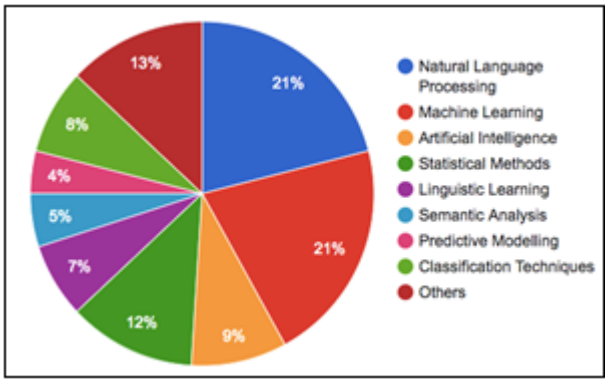


**Figure 2**: Popular Text mining Technologies [2]

### 2.2. Machine Learning and Artificial Intelligence (ML and AI)

ML techniques like clustering and classification models are used to extract meaningful insights and from the vast ocean of data provided. AI techniques like neural networks help in developing predictive and pattern recognition models for semantic analysis and opinion mining.

### 2.3. Statistical Methods and Predictive Modeling

Statistical methods and predictive modeling constitute substantial percent of the total methods used by the popular tools.

### 2.4. Linguistic Learning

This analysis approach is based on the knowledge about the grammar, ontologies and semantics of the language. Linguistic learning learns the structure, the morphology of the language to understand negation and conditionality of the textual data. This technique thus gives an actual and a detailed representation of the structure of the sentence.

### 2.5. TopicModeling

It is often difficult to parse through large documents in limited time. In such cases, Topic modeling is used to summarize and organize large documents based on the topics and categories that best describe the underlying idea of the entire document. This technique uses probabilistic topic model Latent Dirichlet allocation for allowing sets of observations to be explained based on the similarity of their occurrence.

### 2.6. Genetic Algorithms

Genetic algorithms are adaptive, evolutionary algorithms that work in an iterative manner. Every string is encoded in binary, real etc. An evaluation function is used as a fitness measure. It is used for web mining, xml mining and opinion mining to produce optimized solution. It applies 'survival of the fittest' in feature selection.

### 2.7 Classification techniques

Classification techniques like kNN, Naive Bayesian, Support Vector Machine, Decision Trees and Regression are used in most of the tools.

## 3. Text Analysis Tools

Due to upcoming advances in the field of big data and data analytics, a lot of tools have been developed to meet the ever-increasing needs of the industry fro the same. From the vast sea of tools made available for the user's disposal, in this paper we have selected one type from each category. The category of tools can be given as :

- Proprietary Tools
- Open Source Tools (Like frameworks)
- Online Tools (Like APIs)

### 3.1. Buzzlogix

Buzzlogix is a text analysis API that allows developers to integrate the services into their native applications, thus enabling the user to understand their customers and predict consumer behavior. Buzzlogix is a Semantic Text Analysis API that is based on Natural Language Processing and Machine Learning.

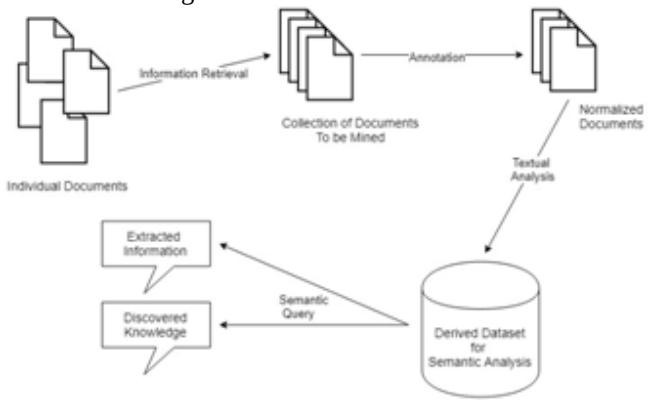


**Figure 3:** Architecture of Semantic Text Analysis

Figure 3 depicts the basic architecture of Semantic Text Analysis. Semantic Text Analysis describes the process of understanding natural language, that is the way humans communicate with each other based on meaning and context [4]. The semantic analysis of any natural language content starts by reading all the words in the content to capture the real meaning of the text. The text element is identified and

assigned a logical and grammatical role. The content in the surrounding text is identified and analyzed with the text structure to accurately disambiguate the proper meaning of words that have more than one interpretation.

Semantic Text Analysis processes a sentence's logical structure to identify the most relevant elements in a text and understand it's context.[4] It also understands the relationships between the different concepts discussed. For example, it understands that a text is about "education" or "schooling" even if it doesn't contain the actual words but related topics such as "grades", "subjects", "exams" or "teachers". Semantic Analysis and Natural Language Processing can help machines automatically understand text, it also supports the larger goal of translating information that is potentially valuable - like customer feedback or an insightful tweet - in the world of business intelligence for customer support service and knowledge management.

### 3.2. Lexalytics

Lexalytics Inc. is a company that provides proprietary software for intent and sentiment analysis using SaaS and cloud based technology.

The intelligence platform provided by Lexalytics is a complete, modular data analytics solution for translating unstructured text documents into valuable information that leads to profitable decisions. Lexalytics works on the principles of Natural Language Processing (NLP).

The biggest challenge in developing traditional NLP applications is that they require humans to "speak" to them in a programming language that is precise, highly structured and unambiguous. Our speech, however, is none of these things. It is often ambiguous and the linguistic structure depends on many complex variables that include slang, regional dialects and social contexts. Current approaches to NLP are based on deep learning, that examines and uses a pattern in a data to improve its understanding. Figure 4 shows the basic steps of NLP.

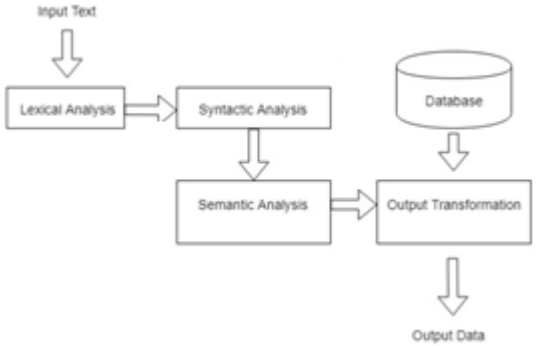


**Figure 4:** Architecture of Natural Language Processing

NLP systems also have a lexicon (a vocabulary) and a set of grammar rules coded into the system. Modern NLP algorithms use statistical machine learning to apply these rules to the natural language and determine the most likely meaning behind what was said[5].

By the end of the processing, the computer understands the meaning of what we said. There are several challenges in accomplishing this when considering problems such as words having several meanings (polysemy) or different words having similar meanings (synonymy), but developers encode rules into their NLP systems and train them to learn to apply the rules correctly.

### 3.3. Microsoft Distributed Machine Learning Toolkit

The Distributed Machine Learning ToolKit is an open-source multi-verso machine-learning framework made available to the developers to train their models with humongous amounts of data. Under this open source project, a number of tools have been made available to the users. These tools include the basic ML library, LightLDA, distributed word embedding and multisense word embedding. The network for word embedding consists of an input layer, embedding layer, hidden layer and an output layer. The word embedding weights are used to map each index to a distributed feature. The architecture for word embedding is shown in Figure 5.

NLP traditionally uses atomic symbols for representing words, which does not provide any semantic relationship among these words thus proving to be of very little use. However, this tool uses the word embedding technique wherein each word in the document is represented in the form of the vector, which is in turn mapped from a vocabulary [6]. The vector representation is such that semantically related words appear to be embedded closer to each other. Thus, we can easily train the statistical models based on the count of how often the words co-occur in the vicinity of each other. Distributed trainer for word2vec is used and the skip-gram model is used as a reference for distributed multi-sense embedding.

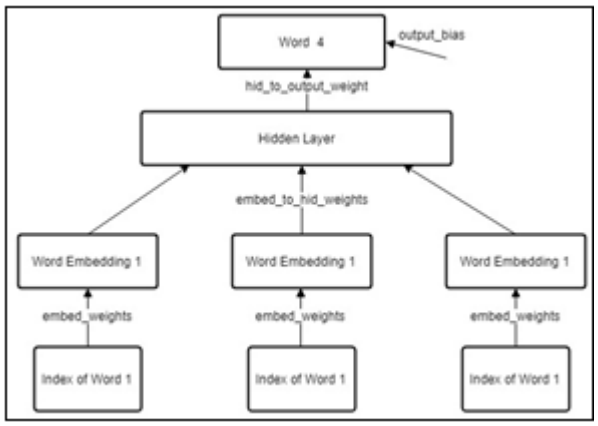


**Figure 5:** Architecture of Word Embedding Network

Word2vec algorithm tries to predict target words from the source words or vice-versa. A dataset of words is created along with the context in which it appears. Context can be syntactic, words appearing on the left of the target word or the words appearing on the right of the target word. The function is defined over this dataset and optimization using the gradient descent is carried out using one example at a time.

This tool also makes use of the Latent Dirichlet Algorithm (LDA) text-mining algorithm for topic modeling [7]. Modeling of discrete data proves to be a time-consuming

and arduous task. In order to make this easy, short descriptions are assigned to the data while the statistical relationship between this data remains untouched. Data is thus modeled into a set of topics and further modeled into topical probabilities. These probabilities provide an acute representation of the entire document in terms of numbers and statistics. LDA is represented in three levels i.e. corpus level, document level and word level [7].

Data parallelism is highly supported on this platform. Data can be parallely trained on different machines and parameters can be updated synchronously or asynchronously.

## 4. Important Features of the Tools

We analyzed the features of tools mentioned in Table1 and the various uses of Text Mining tools[1]. The major uses of a text-mining tool are for:

### 4.1. Text Analytics:

Text analysis involves the extraction of useful information and patterns from unstructured text.

### 4.2. Text Processing

Text processing involves the transformation and manipulation of unstructured text so that different analysis methods can be applied to it.

### 4.3.Classification/Categorization:

Many tools are used for classification and categorization of text based on their content

### 4.4. Sentiment Analysis

Sentiment analysis is used to identify subjective information from any text and helps understand the meaning of the text. It is also called as Opinion Mining.

### 4.5.Knowledge Discovery

Knowledge discovery deals with identification of useful information from large chunks of text.

### 4.6. Semantic Analysis

Sentiment Analysis involves checking the syntactic structures with the meaning of the text as a whole.

## 5. Comparison of Tools

In Table 1, we compare the tools discussed in Section 3 based on their types, the technologies used, the key features and their limitations.

**Table 1:** Comparison of Text Mining Tools

| *Property* | *Buzzlogix* | *Lexalytics* | *Microsoft DMTK* |
|---|---|---|---|
| Type | API | Cloud Based Tool | Open Source framework |
| Technology | Semantic Text Analysis | Natural Language Processing | Word Embedding |
| Capabilities | Text analysis, Semantic analysis, classification, keyword analysis | Sentiment Analysis, Categorization, Named Entity Extraction | Text and Big Data analysis |
| Features | Text Sentiment Analysis, Twitter sentiment analysis, subjective analysis, keyword extraction, entity extraction, classification and more | Cloud Based, easy access, different business packages for different users | Open Source, underlying machine learning framework, distributed multi-sense word embeddings carried out over the platform, topic modeling, data parallelization |
| Limitations | Proprietary, needs SDK | Proprietary | SDK needed |

## 6. Conclusion

Text Analysis is a new and upcoming field of research and there are many tools available for text analysis and text mining. But handling unstructured data is still a challenge. Since most data these days is in the form of text i.e. unstructured, there is a need for text mining and analysis to be much more efficient. Almost all organizations have to deal with textual data, like online articles, Facebook posts or even Tweets. This data can be used to formulate business strategies and make key business decisions. Text Analysis is a growing field with an ever-increasing scope. The market of text analysis tools has changed multifold in the past years and with the advent of intelligent technology and efficient algorithms, it is only going to expand.

## References

[1] Kaur, Arvinder& Chopra, Deepti. (2016). Comparison of text mining tools. 186-192. 10.1109/ICRITO.2016.7784950.
[2] Tan, Ah-Hwee& Ridge, Kent & Labs, Digital &MuiKeng Terrace, Heng. (2000). Text Mining: The state of the art and the challenges.
[3] Yang, Yunyun& Akers, Lucy &Klose, Thomas & Yang, Cynthia. (2008). Text mining and visualization tools-Impressions of emerging capabilities.World Patent Information. 30. 280-293. 10.1016/j.wpi.2008.01.007.
[4] Liang-Yan Li, Zhong-Shi He and Yong Yi, "Principles and algorithms of semantic analysis," *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.03EX693)*, Xi'an, 2003, pp. 1613-1618 Vol.3. doi: 10.1109/ICMLC.2003.1259754
[5] T. Patten and P. Jacobs, "Natural-language processing," in *IEEE Expert*, vol. 9, no. 1, pp. 35-, Feb. 1994. doi: 10.1109/64.295134

[6] Li, Yang & Yang, Tao. (2017). Word Embedding for Understanding Natural Language: A Survey. 26. 10.1007/978-3-319-53817-4.
[7] Tong, Zhou & Zhang, Haiyi. (2016). A Text Mining Research Based on LDA Topic Modelling.Computer Science & Information Technology. 6. 201-210. 10.5121/csit.2016.60616.