# Abstract

**Text mining** refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modelling (*i.e.*, learning relations between named entities).

Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

In our project we build a text mining module which has sub modules for sentiment analysis, gender identification from names, document summarization, and topic recognition using DBpedia.

For sentiment analysis we build a hybrid model which uses two approaches – lexicon based approach and machine learning approach. For gender identification we use a machine learning approach where we train a Naïve Bayes classifier to classify names. The summarization technique makes use of Luhn's summarization algorithm. It basically rests on the premise that the most important sentences are a good summary of the content if presented in chronological order, and that you can discover the most important sentences by identifying frequently occurring words that interact with one another in close proximity. Although a bit crude, this form of summarization works surprisingly well on reasonably well-written Web content. For topic recognition we describe a method for identifying topics in text published in social media, by applying topic recognition techniques that exploit DBpedia.