Natural Language Processing (NLP)

Text Analytics (TA)

Knowledge-Discovery in Text (KDT)

Named entity recognition (NER)

Hypertext markup language HTML

Information extraction (IE)

ATLAS-

WWW- World Wide Web

Db- Data Base

POS- point of

C-Link- concept linkage

K-average- knowledge average

Question Answering Question answering (Q&A)

Recurrent Neural Networks (RNN)

Convolutional Neural Networks (CNN)

APIs-

Social Media Analytics (SMA)

RSS-

IBM

RL- Reinforcement learning

emotional quotient (EQ)

Instituto de Ingeniería del Conocimiento (IIC)

Abstract

The analysis of the text content in emails, blogs, tweets, forums and other forms of textual communication constitutes what we call text analytics. Text analytics is applicable to most industries: it can help analyze millions of emails; you can analyze customers’ comments and questions in forums; you 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, Named Entity Recognition, Semantic Web annotated domain’s representation, 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.

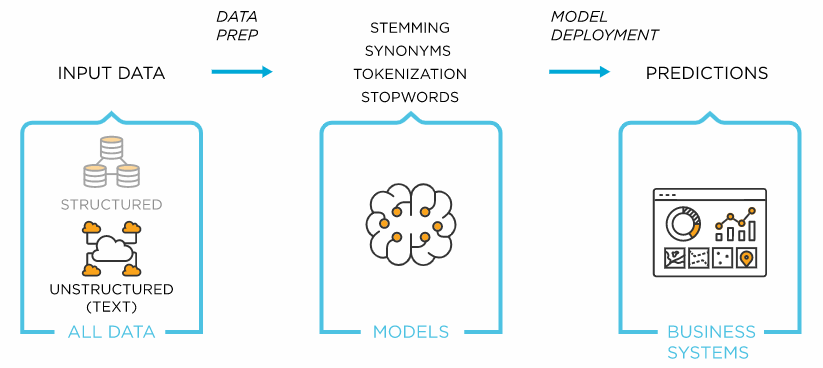
Chapter One

1. Introduction

Natural Language Processing (NLP) is the practical field of Computational Linguistics, although some authors use the terms almost interchangeably. Sometimes NLP has been considered a sub discipline of Artificial Intelligence, and more recently it sits at the 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: Natural Language Understanding, Natural Language Generation, Knowledge Base building, 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 the most recent name given to Natural Language Understanding, Data and Text Mining. In the last few years a new name has gained popularity, Big Data, to refer mainly to unstructured text (or other information sources), more often in the commercial rather than the academic area, probably because unstructured free text accounts for 80% in a business context, including tweets, blogs, wikis and surveys . In fact there is a lack of academic papers covering this topic, although this may be changing in the near future. 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: Concepts and Techniques 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, also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT), refers generally to the process of extracting non-trivial information and knowledge from unstructured text. Text Analytics is similar to data mining, except that data mining tools are designed to handle structured data from databases, either stored as such or as a result from preprocessing unstructured data. Text Analytics can cover unstructured or semi-structured data sets such as emails, full-text documents and HTML files, blogs, newspaper articles, academic papers, etc. Text Analytics is an interdisciplinary field which draws on information extraction, data mining, machine learning, statistics and computational linguistics. Text Analytics is gaining prominence in many industries, from marketing to finance, because the process of extracting and analyzing 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, Bio Medicine or Finance, data scientists, psychologists, causing that the research and development approach is fragmented due to different traditions, methodologies and interests.

Definition of text analytics:-

Text analytics combines a set of [machine learning](https://www.tibco.com/reference-center/what-is-machine-learning), statistical and linguistic techniques to process large volumes of unstructured text or text that does not have a predefined format, to derive insights and patterns. It enables businesses, governments, researchers, and media to exploit the enormous content at their disposal for making crucial decisions. Text analytics uses a variety of techniques – sentiment analysis, topic modelling, named entity recognition, term frequency, and event extraction.



What’s the Difference between Text Mining and Text Analytics?

Text mining and text analytics are often used interchangeably. The term text mining is generally used to derive qualitative insights from unstructured text, while text analytics provides quantitative results.

For example, text mining can be used to identify if customers are satisfied with a product by analyzing their reviews and surveys. Text analytics is used for deeper insights, like identifying a pattern or trend from the unstructured text. For example, text analytics can be used to understand a negative spike in the customer experience or popularity of a product.

The results of text analytics can then be used with [data visualization techniques](https://www.tibco.com/reference-center/guide-to-data-visualization) for easier understanding and prompt decision making.

**Chapter two**

**Applications of text analytics**

**A** typical text analytics application consists of the following steps and tasks: Starting with a collection of documents, a text mining tool retrieves a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. The underlying strategy in all the components is to find a pattern (from either a list or a previous process) which matches a rule, and then to apply the rule which annotates the text. Each component performs a particular process on the text, such as: sentence segmentation (dividing text into sentences); tokenization (words identified by spaces between them); part-of-speech tagging (noun, verb, adjective, etc., determined by look-up and relationships among words); shallow syntactic parsing/ chunking (dividing the text by noun phrase, verb phrase, subordinate clause, etc.); named entity recognition (NER) (the entities in the text such as organizations, people, and places); dependency analysis (subordinate clauses, pronominal anaphora [i.e., identifying what a pronoun refers to], etc.). The resulting process provides “structured” or semi-structured information to be further used (e.g. Knowledge Base building, Ontology enrichment, Machine Learning algorithm validation, Query Indexes for Question & Answer systems). Some of the techniques that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, question answering, and deep learning. International Journal of Interactive Multimedia and Artificial Intelligence.

1. Information Extraction

Information extraction (IE) software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process usually called pattern matching, typically based on regular expressions. 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. These are several tools relevant for this task: Apache Open NLP, Stanford Named Entity Recognizer.

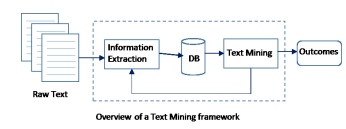


Fig. 1. Overview of a Text Mining Framework

1. Topic Tracking and 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. Manual keyword extraction is an extremely difficult and time consuming task; in fact, it is almost impossible to extract keywords manually in case of news articles published in a single day due to their volume. A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Google offers a free topic tracking tool that allows users to choose keywords and notifies them when news relating to those topics becomes available. NER techniques are also used in enhancing topic tracking and detection by matching names, locations or usual terms in a given topic by representing similarities with other documents of similar content.

C. Summarization

Text 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. With large texts, text summarization processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. One of the strategies most widely used by text summarization tools is sentence extraction. Important sentences from an article are statistically weighted and ranked. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. The methods of summarization can be classified in two broad groups: • shallow analysis, restricted to the syntactic level of representation and try to extract significant parts of the text; • deeper analysis, assumes a semantics level of representation of the original text (typically using Information Retrieval techniques). A relatively recent European Union project, ATLAS, has performed an extensive evaluation of text summarization tools.

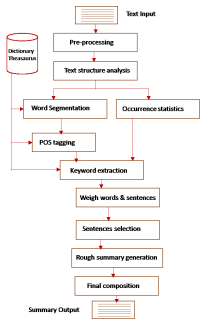
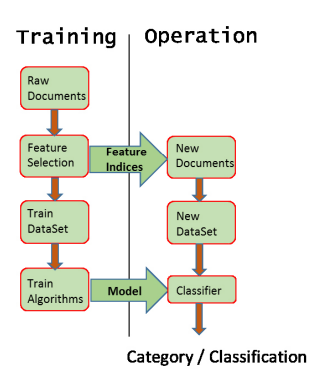


Fig. 2. Text Summarization

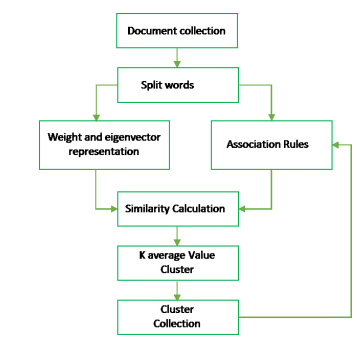
1. Categorization or Classification

Categorization involves identifying the main themes of a document by placing the document into a predefined set of topics (either as taxonomies or ontologies). 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. 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.



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. Using clustering, documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results (multiple indexing references). A basic clustering algorithm creates a vector of topics for each document and assigns the document to a given topic cluster. Fig. 4. Document Clustering Medicine and Legal research papers have been a fertile ground to apply text clustering techniques.



E. Concept Linkage

Concept linkage tools connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps would not have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical and legal fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research. The best known concept linkage tool is C-Link. C-Link is a search tool for finding related and possibly unknown concepts that lie on a path between two known concepts. The tool searches semi structured information in knowledge repositories based on finding previously unknown concepts that lie between other concepts.

F. Information Visualization

Visual text mining, or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics. A common typical example of text information visualization are Tag clouds, like those provided by tools such as Wordle. Hearst has written an extensive overview of current (and recent past) tools for text mining visualization, but definitively needs an update with the appearance of new tools in recent years: Fig. 5. Text data visualization (Source: Hearst (2009) “Information Visualization for Text Analysis”)

G. Question Answering Question answering (Q&A) systems used natural language queries to find the best answer to a given question. Question answering involves a lot of techniques described here, from information extraction for the question topic understanding, question typology and categorization, up to the actual selection and generation of the answer

H. Deep Learning Deep Learning has been gaining a lot of popularity as of the last two years, and has begun to be experimented for some NLP tasks. Deep Learning is a very broad field and most promising work is moving around Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). Neural Networks have a long and prestigious history, and interest within the field of Text Analytics has been revived recently. In a traditional neural network all inputs (and outputs) are independent of each other. The idea behind RNNs is to make use of sequential information (as the words in a sentence). In order to predict the next word in a sentence we must know which words came before it. RNNs perform the same task for every element of a sequence, with the output being dependent on the previous computations. RNNs have a “memory” which captures information about what has been calculated so far. With RNN a given language model can be built, which in turn allows to score arbitrary sentences based on how likely they are to occur in the real world, and later that model allows to generate new text. CNNs are basically just several layers of convolutions over the input layer to compute the output. Each layer applies different filters, typically hundreds or thousands, and combines their results. These can be used for Sentiment Analysis, Spam Detection or Topic Categorization, but they are of little use with POS Tagging or Entity Extraction unless additional features are included as filters. DL4J is a tool for textual analysis using deep learning techniques. It builds on Word2vec, a two-layer neural network that processes text, created by Google.

III. Known Problems in Text Analytics In the context of TA, Big Data is simply a massive volume of written language data. But where does the frontier lie between Big Data and Small Data? There has been a culture-changing fact: while merely 15 years ago a text corpus of 150 million words was considered huge, currently no less than 8.000 million word datasets are available. Not only is it a question simply about size, but also about quality and veracity: data from social media are full of noise and distortion. All datasets have these problems but they are more potentially serious for large datasets because the computer is an intermediary and the human expert do not see them directly, as is the case in small datasets. Therefore, data cleansing processes consume significant efforts and often after the cleansing, the availability of information to train systems is not enough to get reliable predictions, as happened in the Google Flu Trends failed experiment. The reason is that most big datasets are not the output of instruments designed to produce valid and reliable data for analysis, and also because data cleansing is about (mostly subjective) decisions on the relevant design features. Another key issue is the access to the data. In most cases, the academic groups have no access to data from companies such as Google, Twitter or Facebook. For instance, Twitter only makes a very small fraction of its tweets available to the public through its APIs. Additionally, the tweets available do not follow a given pattern (they are an “assorted bunch”) so it is difficult to arrive at a conclusion concerning their representativeness. As a consequence, the replication of analyses is almost impossible, since the supporting materials and the underlying technology are not publicly available. Boyd and Crawford go further: limited access to Big Data creates new digital divides, the Big Data rich and the Big Data poor. One needs the means to collect them, and the expertise to analyze them. Interestingly, small but well curated collections of language data (the traditional corpora) offer information that cannot be inferred from big datasets. How to grasp the figurative uses of language, basically irony and metaphor, is also a well-known problem to properly understand text. Essentially, the user’s intentions are hidden because the surface meaning is different to the underlying meaning. As a consequence, the words must be interpreted in context and with extra-linguistic knowledge, a fact that being hard on humans, it is even harder for machines. How to translate a given metaphor into another language is extremely difficult. Some estimates calculate that figurative language is about 15-20% of the total content in social media conversations. IV. Some Use Cases Text Analytics has produced useful applications for everyday use.

Contextual comprehension – Concepts and ideas in the human brain are semantically linked. Amelia can quickly and reliably retrieve information across a wider and more complex set of knowledge. • Emotional responsiveness – Research shows that a better user experience is directly tied to empathy shown by the agent throughout the interaction. In addition to an increased emotional quotient (EQ), Amelia presents a mood and a personality vector in a 3-dimensional emotional space. The software is already used for services such as technology helpdesks, contact centers, procurement processing and to advise field engineers, among other business processes. Both Watson and Amelia were already briefly mentioned in Redondo, as the purpose of both systems is to extend a human’s capabilities by applying intelligent artificial systems techniques, such as deep learning and interpersonal communication through dialogue management.

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 Legal Analytics – Information extraction from legal texts.

1. 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 (like Unified Medical Language System-UMLS, or SNOMED-CT) 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). The NLP techniques include biomedical entities recognition, pattern recognition, and machine learning for extracting semantic relations between concepts. Biomedical entities recognition consists of recognizing and categorizing entity names in biomedical domains, such as proteins, genes, diseases, drugs, organs and medical specialties.

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. They are typically based on lexical resources and their main drawbacks are the abundance of synonyms and spelling variations for entities and concepts.

• 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. Their main shortcoming is the requirement of computationally expensive training and testing on large amounts of human-tagged data. To extend the extraction system to another type of data or language requires new human effort in annotation. Friedman et al. [42] presents a survey of the state of the art and prospects in BioNLP, sponsored by the US National Library of Medicine. This report identifies that “the most significant confounding element for clinical NLP is inaccessibility of large scale de-identified clinical corpora, which are needed for training and evaluation.” 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. Processing Natural Language (NL) to support such richly annotated documents presents some inherent issues. NL supports all of the following, among other things:

(1)implicit or presupposed information – “When did you stop taking drugs?” (presupposes that the person is questioned about taking drugs at some time in the past);

(2)multiple forms with the same meaning – Jane Smith, Jane R. Smith, Smith, Attorney Smith… (our NER system must know that these are different ways to refer to the same physical person);

(3)the same form with different contextually dependent meanings – An individual referred to as “Jane Smith” in one case decision may not be the individual referred to by the name “Jane Smith” in another case decision; and

(4) Dispersed meanings – Jane Smith represented Jones Inc. She works for Boies, Schiller and Flexner. To contact her, write to j.smith@bsfllp.com People grasp naturally relationships between words and phrases, such that if it is true that Bill used a knife to injure Phil, then Bill injured Phil. Natural Language Processing (NLP) addresses this highly complex problem as an engineering problem, decomposing large problems into smaller problems and subdomains (e.g. summarization, information extraction, …) that we have already covered above. In the area of Legal Analytics we are primarily interested in information extraction. Typically a legal analytics system will annotate elements of interest, in order to identify a range of particular pieces of information that would be relevant to legal professionals such as:

• Case citation

• Names of parties

• Roles of parties, meaning plaintiff or defendant

• Type of court

• Names of judges

• Names of attorneys

• Roles of attorneys, meaning the side they represent (plaintiff or defendant)

• Final decision

• Cases cited

• Nature of the case, meaning using keywords to classify the case in terms of subject (e.g., criminal assault, intellectual property, etc.) The business implications of Legal Analytics have originated a full branch of textual Big Data applications.).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 emojis 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.

Chapter Three

## What’s the Relevance of Text Analytics in Today’s World?

As of 2020, around 4.57 billion people have access to the internet. That’s roughly 59 percent of the world’s population. Out of which, about 49 percent of people are active on social media. An enormous amount of text data is generated every day in the form of blogs, tweets, reviews, forum discussions, and surveys. Besides, most customer interactions are now digital, which creates another huge text database.

Most of the text data is unstructured and scattered around the web. If this text data is gathered, collated, structured, and analyzed correctly, valuable knowledge can be derived from it. Organizations can use these insights to take actions that enhance profitability, customer satisfaction, research, and even national security.

## Benefits of Text Analytics

There are a range of ways that text analytics can help businesses, organizations, and event social movements:

* Help businesses to understand customer trends, product performance, and service quality. This results in quick decision making, enhancing [business intelligence](https://www.tibco.com/reference-center/what-is-business-intelligence), increased productivity, and cost savings.
* Helps researchers to explore a great deal of pre-existing literature in a short time, extracting what is relevant to their study. This helps in quicker scientific breakthroughs.
* Assists in understanding general trends and opinions in the society, that enable governments and political bodies in decision making.
* Text analytic techniques help search engines and information retrieval systems to improve their performance, thereby providing fast user experiences.
* Refine user content recommendation systems by categorizing related content.

## Text Analytics Techniques and Use Cases

There are several techniques related to analyzing the unstructured text. Each of these techniques is used for different use case scenarios.

### Sentiment analysis

Sentiment analysis is used to identify the emotions conveyed by the unstructured text. The input text includes product reviews, customer interactions, social media posts, forum discussions, or blogs. There are different types of sentiment analysis. Polarity analysis is used to identify if the text expresses positive or negative sentiment. The categorization technique is used for a more fine-grained analysis of emotions - confused, disappointed, or angry.

Use cases of sentiment analysis:

* Measure customer response to a product or a service
* Understand audience trends towards a brand
* Understand new trends in consumer space
* Prioritize customer service issues based on the severity
* Track how customer sentiment evolves over time

### Topic modelling

This technique is used to find the major themes or topics in a massive volume of text or a set of documents. Topic modeling identifies the keywords used in text to identify the subject of the article.

Use cases of topic modeling:

* Large law firms use topic modeling to examine hundreds of documents during large litigations.
* Online media uses topic modeling to pick up trending topics across the web.
* Researchers use topic modeling for exploratory literature review.
* Businesses can determine which of their products are successful.
* Topic modeling helps anthropologists to determine the emergent issues and trends in a society based on the content people share on the web.

### Named Entity Recognition (NER)

NER is a text analytics technique used for identifying named entities like people, places, organizations, and events in unstructured text. NER extracts nouns from the text and determines the values of these nouns.

Use cases of named entity recognition:

* NER is used to classify news content based on people, places, and organizations featured in them.
* Search and recommendation engines use NER for information retrieval.
* For large chain companies, NER is used to sort customer service requests and assign them to a specific city, or outlet.
* Hospitals can use NER to automate the analysis of lab reports.

### Term frequency – inverse document frequency

TF-IDF is used to determine how often a term appears in a large text or group of documents and therefore that term’s importance to the document. This technique uses an inverse document frequency factor to filter out frequently occurring yet non-insightful words, articles, propositions, and conjunctions.

### Event extraction

This is a text analytics technique that is an advancement over the named entity extraction. Event extraction recognizes events mentioned in text content, for example, mergers, acquisitions, political moves, or important meetings. Event extraction requires an advanced understanding of the semantics of text content. Advanced algorithms strive to recognize not only events but the venue, participants, date, and time wherever applicable. Event extraction is a beneficial technique that has multiple uses across fields.

Use cases of event extraction:

* Link analysis: This is a technique to understand “who met whom and when” through event extraction from communication over social media. This is used by law enforcement agencies to predict possible threats to national security.
* Geospatial analysis: When events are extracted along with their locations, the insights can be used to overlay them on a map. This is helpful in the geospatial analysis of the events.
* Business risk monitoring: Large organizations deal with multiple partner companies and suppliers. Event extraction techniques allow businesses to monitor the web to find out if any of their partners, like suppliers or vendors, are dealing with adverse events like lawsuits or bankruptcy.

## Steps Involved with Text Analytics

Text analytics is a sophisticated technique that involves several pre-steps to gather and cleanse the unstructured text. There are different ways in which text analytics can be performed. This is an example of a model workflow.

1. Data gathering - Text data is often scattered around the internal databases of an organization, including in customer chats, emails, product reviews, service tickets and Net Promoter Score surveys. Users also generate external data in the form of blog posts, news, reviews, social media posts and web forum discussions. While the internal data is readily available for analytics, the external data needs to be gathered.
2. Preparation of data - Once the unstructured text data is available, it needs to go through several preparatory steps before machine learning algorithms can analyze it. In most of the text analytics software, this step happens automatically. Text preparation includes several techniques using natural language processing as follows:
   * Tokenization: In this step, the text analysis algorithms break the continuous string of text data into tokens or smaller units that make up entire words or phrases. For instance, character tokens could be each individual letter in this word: F-I-S-H. Or, you can break up by subword tokens: Fish-ing. Tokens represent the basis of all natural language processing. This step also discards all the unwanted contents of the text, including white spaces.
   * Part-of-speech-tagging: In this step, each token in the data is assigned a grammatical category like noun, verb, adjective, and adverb.
   * Parsing: Parsing is the process of understanding the syntactical structure of the text. Dependency parsing and constituency parsing are two popular techniques used to derive syntactical structure.
   * Lemmatization and stemming: These are two processes used in data preparation to remove the suffixes and affixes associated with the tokens and retain its dictionary form or lemma.
   * Stopword removal: This is the phase when all the tokens that have frequent occurrence but bear no value in the text analytics. This includes words such as ‘and’, ‘the’ and ‘a’.
3. Text analytics - After the preparation of unstructured text data, text analytics techniques can now be performed to derive insights. There are several techniques used for text analytics. Prominent among them are text classification and text extraction.

Text classification: This technique is also known as text categorization or tagging. In this step, certain tags are assigned to the text based on its meaning. For example, while analyzing customer reviews, tags like “positive” or “negative” are assigned. Text classification often is done using rule-based systems or machine learning-based systems. In rule-based systems, humans define the association between language pattern and a tag. “Good” may indicate positive review; “bad” may idenitfy a negative review.

Machine learning systems use past examples or training data to assign tags to a new set of data. The training data and its volume are crucial, as larger sets of data helps the machine learning algorithms to give accurate tagging results. The main algorithms used in text classification are Support Vector Machines (SVM), Naive Bayes family of algorithms (NB), and deep learning algorithms.

Text extraction: This is the process of extracting recognizable and structured information from the unstructured input text. This information includes keywords, names of people, places and events. One of the simple methods for text extraction is regular expressions. However, this is a complicated method to maintain when the complexity of input data increases. Conditional Random Fields (CRF) is a statistical method used in text extraction. CRF is a sophisticated but effective way of extracting vital information from the unstructured text.

.

Chapter three

Conclusion

Text Analytics, with its long and prestigious history, is an area in constant evolution. It sits at the center 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.

## What Happens After Text Analytics? Once the text analytics methods are used to process the unstructured data, the output information can be fed to data visualization systems. The results can then be visualized in the form of charts, plots, tables, infographics, or dashboards. This visual data enables businesses to quickly spot trends in the data and make decisions

Reference

<https://www.tibco.com/reference-center/what-is-text-analytics>

https://reunir.unir.net/bitstream/handle/123456789/11218/ijimai20163\_6\_9\_pdf\_38545.pdf?sequence=1&isAllowed=y