

**UNIVERSITY OF GONDAR**

**FACULITY OF INFORMATICS**

**DEPARTMENT OF INFORMATION TECHNOLOGY FOR**

**3rd YEAR SUMMER PG PROGRAM**

Seminar

Assignment

Submitted To:- Newayneh K. (PhD)

Submission Date: 20/12/2014 E.C

Prepared by:

1. Belay Eyasu Damena Gus/00144/11
2. Adane Worku Gelaye Gus/00128/11
3. Belayneh Ayalew Mekonen Gus/00150/11
4. Delelegn Desie Fentie Gus/00010/10
5. Endayen Berihun Sisay Gus /22336/10

Gondar, Ethiopia

August 2014E.

# CERTIFICATION

# This is to certify that Seminar assignment was carried out by the following group members

# Group name ID Signature Date

1. Belay Eyasu Damena Gus/00144/11 \_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_
2. Adane Worku Gelaye Gus/00128/11 \_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_
3. Belayneh Ayalew Mekonen Gus/00150/11 \_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_
4. Delelegn Desie Fentie Gus/00010/10 \_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_
5. Endayen Berihun Sisay Gus /22336/10 \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_

Approved by :- **Tsehay W(Phdc) Signature \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Date \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Acknowledgment

Any accomplishment requires the blessing of the God, So thanks to God. Next we would like to thank our honorable Instructor **Tsehay (Phdc)** for his continuous suggestions in every ways how to prepare the seminar assignment and give so much enough time to do the assignment.

Finally we also would like to thank university of university of Gondar digital library staffs for accomplishment of this assignment

# Table of content

# List of figure

# Acronyms

**C-Link**- concept linkage

**CNN**-Convolutional Neural Networks

**DB**- Data Base

**EQ**-emotional quotient

**HTML**- Hypertext markup language

**IBM**- International Business machine

**IE**-Information extraction

**IIC**-Institute de Ingeniería del Conocimiento

**K-average**- knowledge average

**KDT**- Knowledge-Discovery in Text

**NER**-Named entity recognition

**NLP**-Natural Language Processing

**POS** -Part of Speech tagging

**Q&A**-Question Answering Question answering

**RL**- Reinforcement learning

**RNN**- Recurrent Neural Networks

**SMA**-Social Media Analytics

**TA**-Text Analytics

**WWW**- World Wide Web

# 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 combines a set of 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.*

*Many companies use text analytics to analyze articles, tweets, social media posts, reviews, comments, and other types of writing, to find meaning and gather intelligence with the help of algorithms and tools.*

*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.*

# 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. 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.

# CHAPTER TWO

# 2. TEXT ANALYTICS

## 2. 1 WHAT IS TEXT ANALYTICS?

Text analytics is an automated process to analyze a piece of writing and extract useful information from it. It is often carried out with the help of software designed to go through lengthy texts and gather insights that may be useful for marketing, branding, and other research purposes.

Many companies use text analytics to analyze articles, tweets, social media posts, reviews, comments, and other types of writing, to find meaning and gather intelligence with the help of algorithms and tools.

[machine learning](https://en.wikipedia.org/wiki/Machine_learning) ..

Text analytics combines a set of 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.

## 2.1 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.

Text mining is very often used as a synonym of text analytics, so these two terms mostly refer to the same concept. However, text mining is a broader term that refers to the act of gathering useful, high-quality information from a text. Instead, text analytics is the more specific computational process of analyzing a text to extract such information. Text analytics software’s use linguistic, statistical, and machine learning techniques to structure the content of a text and analyze it. In order to perform such analysis, programs use natural language processing (commonly abbreviated as NLP), which is the field of computer science, information engineering, and artificial intelligence used to process and analyze the data contained in a piece of writing.

## 2.3 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. 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.

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).

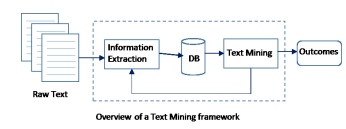


Fig. 1. Overview of a Text Mining Framework

1. **Topic Tracking and Detection**

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.

The methods of summarization can be classified in two broad groups:

A relatively recent European Union project has performed an extensive evaluation of text summarization tools.

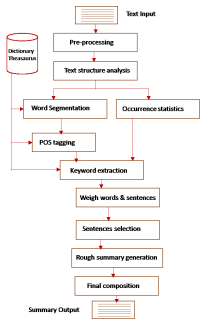
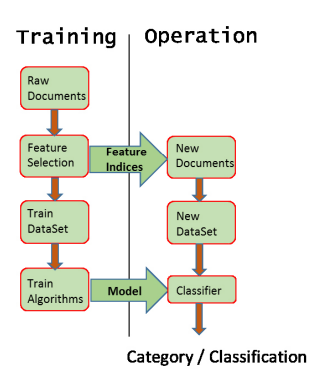


Fig. 2. Text Summarization

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. 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.



1. Fig 3. Categorization or Classification

D. Clustering

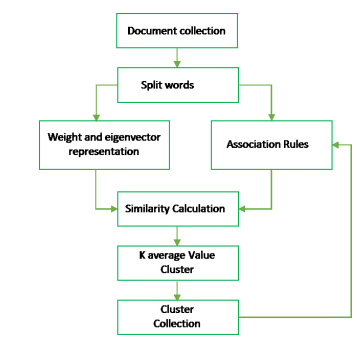
Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered without the use of predefined topics. In other words, while categorization implies supervised (machine) learning in the sense that previous knowledge is used to assign a given 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

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.: G. 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). 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.

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).
2. 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.

## 2.4 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.
* Every day, we create a colossal amount of digital data. In fact, in the [last two years alone we generated 90% percent of all the data in the world](https://techjury.net/blog/how-much-data-is-created-every-day/). That includes social media messages, emails, Google searches, and every other source of online data.
* At the same time, books, media libraries, reports, and other types of databases are now available in digital format, providing researchers of all disciplines opportunities that didn’t exist before.
* Text analysis enables businesses to go through massive collections of data with minimum human effort, saving precious time and resources, and allowing people to focus on areas where they can add more value.

## 2.5 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 sub-word 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.

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.

# Chapter four

# 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.

## Finally 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

[machine learning](https://en.wikipedia.org/wiki/Machine_learning)

[data visualization techniques](https://www.tibco.com/reference-center/guide-to-data-visualization)

[business intelligence](https://www.tibco.com/reference-center/what-is-business-intelligence)

* [Word frequency](https://monkeylearn.com/textual-analysis/#word-frequency)
* [Text classification](https://monkeylearn.com/text-classification/)
* [last two years alone we generated 90% percent of all the data in the world](https://techjury.net/blog/how-much-data-is-created-every-day/)

<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