KOCAELI UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER'S THESIS

A USER BASED COMPARATIVE STUDY OF AUTOMATIC TEXT SUMMARIZATION

NAJMA ABDI OMAR

KOCAELI 2018

KOCAELI UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER'S THESIS

A USER BASED COMPARATIVE STUDY OF AUTOMATIC TEXT SUMMARIZATION

NAJMA ABDI OMAR

Prof. Dr. Nevcihan DURU
Supervisor, Kocaeli Univ. .....

Assoc. Prof. Dr. Sevinç İLHAN OMURCA
Jury member, Kocaeli Univ. .....

Prof. Dr. Cemil ÖZ
Jury member, Sakarya Univ. .....

Thesis Defense Date: 11.07.2018

PREFACE AND ACKNOWLEDGEMENT

An average user is usually bombarded with huge data that is difficult to consume and keep up with on a daily bases. With the need to keep up to date with the on goings and the limited time that a user has, there is a direct need for a system that will provide the user with a way to quickly decide which information to keep or which to discard that is where text summarization comes in. Takes summarization takes the input document and provides the user with a shorter version of the original document which could serve as a stand-alone sufficient quick to absorb resource or as a means to decide what information is important from that which is not worthy. In the thesis text summarization has been explored and the user placed as the central importance in an evaluation method proposed.

I would like to express my gratitude to my advisor Prof. Dr. Nevcihan Duru who has introduced me to data and text mining and has guided me throughout both the research and my course work. My sincere thanks to my family and friends who have supported me and believed in me throughout my life journey.

June - 2018 Najma Abdi OMAR

CONTENTS

LIST OF FIGURES

LIST OF TABLES

SYMBOLS AND ABREVIATIONS

HCI	: Human Computer Interaction
LSA	: Latent Semantıc Analysıs
MMR	: Maximal Marginal Relevance
NLP	: Natural Language Processing
NLTK	: Natural Language Tool Kit
RNN	: Recurrent Neural Network
ROUGE	: Recall-Oriented Understudy for Gisting Evaluation
SUS	: System Usability Scale
SVD	: Singular Value Decomposition
TF-IDF	: Term-Frequency-Inverse Document Frequency

OTOMATIK METİN ÖZETLEMEDE KULLANICI TABANLI KARŞILAŞTIRMA ÇALIŞMASI

ÖZET

Her geçen zamanda, metinsel verilerdeki büyüklük devasa ölçülerde olmaktadır. Orta düzeyde bir kullanıcı, normalde işleyebileceğinden daha fazla bilgi ile karşılaşmaktadır. Uzmanlık seviyesinden bağımsız olarak, her kullanıcı kendi ilgi alanına göre, bir şeyler anlatmakta olup, ilaveten teknolojinin gelişmesi ve sosyal medyanın kullanımındaki artışla birlikte, hem sayı hem de boyutta fikir ve haber makalesi sayısı artmıştır. Bu verilerden okurken, değerli malzemelere ulaşma olasılığı daha da azalmakta, dolayısıyla metin özetlemesi gibi teknolojik yeniliklere duyulan ihtiyaç da artmaktadır. Otomatik Metin Özetlemesi, mevcut bilgi miktarını azaltmak için doğal dil işleme ve yapay zekâ gibi alanlarda kullanır. Bu anlamda, otomatik metin özetleme, hangi makalenin daha fazla okunacağına ve hangisinin önemsiz olduğuna karar vermek için bir araç olarak kullanılabilen önemli bir araçtır. Bu tez, dört farklı algoritmanın karşılaştırmalı bir çalışmasını içermektedir; Gensim TextRank, Sumy LSA, bir frekans olay tabanlı özetleme ve bir duygu analizi tabanlı özetleyici. Çalışmanın değerlendirmesinde, özetlemenin, ortalama bir kullanıcının bir kullanıcı geribildirim anketi yoluyla özetleyicileri puanlama yöntemiyle insan bilgisayar etkileşim çalışması olarak yapılmış ve değerlendirme metrikleri Rouge puanlarıyla karşılaştırılmıştır. Tezde tanımlanan çalışma, otomatik metin özetlemesinin bir arka plan çalışmasını, dört algoritmanın karşılaştırmalı bir incelemesini ve özetleyicilerin değerlendirilmesini içermektedir.

Anahtar Kelimeler: Değerlendirme Yöntemleri, Karşılaştırma Çalışması, Kullanılabilirlik Çalışması, Otomatik Metin Özetleme.

A USER BASED COMPARATIVE STUDY OF AUTOMATIC TEXT SUMMARIZATION

ABSTRACT

Every second that goes by, textual data is generated in magnitudes. The average user is met with more information than they could ordinarily process. With the advent of technology and the rise in the use of social media, opinions and news article extracts have grown in both number and size, every user regardless of expertise level has something to tell the world. In navigating this data, the possibility of getting worthy material is getting slimmer hence the need for technological innovations like text summarization. Automatic Text summarization uses knowledge in the fields like natural language processing and artificial intelligence to downsize on the amount of existing information. It is a great asset that can be used as a tool to decide which article to read further and which one to discard. In this thesis, the work done involves a comparative study of four different algorithm; Gensim TextRank-Based, Sumy LSA-based (borrowed from the original implementation), a frequency event based summarization (simple summarizer) and a sentiment analysis based summarizer. In the evaluation of the study, due to the fact that summarization is centred towards making the work of the average user simpler, a popular Human Computer interaction study was borrowed to score the summarizers through a usability study and a user feedback survey. The evaluation metrics was compared to Rouge scores. The work described in the thesis include a background study of automatic text summarization, a comparative study of the four algorithms and evaluation of the summarizers.

Keywords: Automatic Text Summarization, Comparative Study, Evaluation Methods, Usability Study.

INTRODUCTION

Every second the clock ticks huge data is generated from different sources of social media outlets to be consumed by the average human. [1] states that with the increase in information technologies social media growth has been experienced, in the beginning before social media 'craze' the usage was target for communication purposes but as the usage grew so has it definition, users now use social media as an outlet to voice opinions on matters ranging from personal, religious to political views whether an expert or not. Links to news articles are at the tip of the fingers, just a click and a user is bombarded with more information than an average human brain could consume, analyse and understand. Over time as the size of the resources grew, the availability of consumable information shot up from scarce to too much while the average human ability to read in lifetime remained constant. The growth in the data sizes have no meaning if the user cannot take advantage of what is provided, it will be the same as having no extra information. This multiple outlet of information is good not only for the researchers but also for the average user in the real world, but without the proper tools the information surplus is as good as nothing if it cannot be utilised for knowledge purposes. With this said the need to establish a way for humans to quickly establish what is meaningful and what is to be discarded cannot be stressed enough. Hence technological advances in the field of text mining come in handy. Researches in areas like document retrieval, information extraction and text summarization are all important aspects in data utilization falling in the field of text mining to get meaning from otherwise mixture of rabbles. For the average user this technologies help to get the query needs. Document retrieval brings up the documents related to a user search term, while information extraction gives important snippets related to user query and text summarization gives the content expressed in a document but at a compressed rate giving the user the idea or the concept without the need to read a whole document. Given this a user can decide whether the document is worth the time or not or in some cases get the summary outline only to give all required parts.

If for example we take news articles, when a user wakes up in the morning to keep up with the real world on goings there is a lot of items to read, with a summarizer they could quickly get up to date with the news and go on with the day.

Text summarization as a field was earliest documented in a research done by Luhn where he came with a proposal to get the top most significant sentences in a technical paper, the motivation was to get summaries that had no human bias in it indicating that by using humans summaries could differ from one human summarizer to another depending on a point of view with some times one user having multiple summaries for the same article depending on the emotion or growth in understanding [2]. With computers however, there was a guarantee of 'stable' summaries where following a set of rules a computer will give the same results any time with no bias, the only human input needed was for the initial set up of the program. Although research in text summarization has grown with time and with the advancement in computer in terms of technologies and capacities and also advent of social media and the data it has made possible, advanced methods have been proposed but the idea that was proposed by the grandfather of text summarization still remains as one of the best simple methods that with time have twerked to produce efficient results. Text summarization has moved from the initial frequency based picking of the best sentence scores to recently use of artificial intelligence and deep learning to create almost human brain like summaries. In [3] text summarization is termed as a 'technological art' because "it's a creative and helpful skill that is improving as the years go by with new angles as it grows". Text summarization journey is discussed in detail in this thesis.

Thesis Motivation:

While the main motivation behind text summarization has been to help give solutions to users in curbing with the increase in the amount of data they are forced to read, evaluation of text summarization have veered away from using the people who the tools have been created for. Various forms of evaluation techniques have been implemented when measuring the performance of a summarizer ranging in types from intrinsic evaluations where summaries were tested in terms of text quality or content quality or the extrinsic evaluation which tested how well a given summary performed given a range of task-based problems to deal with while interacting with data. However

in the study of previous research in this area it was noted that most users who summarization targets have been left out of the evaluation process. This aspect formed the strongest motivation in carrying out a usability study to include users in the process of text summarization.

In the field of human computer interaction (HCI) a "multidisciplinary field of study focusing on the design of computer technology and, in particular, the interaction between humans (the users) and computers" [4]. HCI as a field was initially centred only on computers, but with time and the advancement of technology and as more human got hands on experience with technological advancement, HCI has grown to include all aspects of technology in order to encourage the creation of positive experiences in all technological design interactions. In order for HCI to succeed in the target, real world users were targeted for any test that was done in any design with the user feedback given utmost importance. Text Summarization and Text mining in general has previously successfully interacted and borrowed from other field giving magnificent results. It was this that motivated the borrowing from the HCI field in the evaluation section, this thesis sought to find how efficient summaries are in terms of meeting the needs of the user, and using user feedback from survey that was adopted from HCI scaling of websites to score summaries in a comparative study of four methods of summarization. In thesis work we reached out to the users to find out if the summarization result was meeting the correct output finally giving users the wheel to decide the driving force of text summarization.

Thesis Question:

The thesis highlights the existing research work done in text summarization, and in the evaluation section interacts with the users. The main motivation being getting the user feedback on whether or not the summaries give help in decision making in factors such as understanding underlying concept of articles based on summary and also whether to keep or discard the article. The summarizers were based on news article and this formed the main hypothesis of the thesis which was:

Hypothesis: Users find a substantial difference between the two options of either using of summarizers and or digging into information blindly.

To support this hypothesis the thesis question to be answered with the work was formed as

Q1. Do summarizers help in identifying which sources are worth reading from those that are a waste of time with no important substance?

Q2. Do users get the concepts or events behind the articles simply from reading the summary with no prior knowledge of the article in hand or the title of the articles?

In order to implement the required study to answer the questions needed to negate or approve of the hypothesis, four different summarizing algorithms were implemented. The four summarizers that were implemented borrowed the techniques from existing implementations. Two of the summarizers were already in existence as form of python libraries, this two were the Gensim TextRank-Based summarizer and the Sumy LSA-based Summarizer. The other two summarizers were an implementation of the frequency based summarizer where one used a definition of events to choose the words that were used for the frequency calculation giving the terms that are usually associated with events in news article an advantage in terms of sentence score this summarizer is referred to as the simple summarizer. The final summarizer was a hybrid of sentiment analysis and frequency based summarizer where before application of the frequency summarizer, sentiment analysis was applied in the text sentences to gauge the polarity of the sentences and group them according to their polarity i.e. positive and negative group and frequency summarizer applied on the larger group discarding the smaller group as unimportant, this summarizer is referred to as the sentiment analysis-based summarizer.

The evaluation of the system was done using users who interacted with the summaries first answering a list of question and then interacting with the article and giving a survey feedback to score the four summarizers that were used to summarize random news articles from internet sources. The remaining work is discussed in the various chapters as outlined below:

Chapter 1: Contains the background information that is important for the thesis work this include the literature review, the history and journey of text summarization, text mining, natural language processing and machine learning.

Chapter 2: Discusses Automatic Text Summarization touching on the types of text summarization and the existing approaches that are used in sentence extraction.

Chapter 3: Discusses the comparative work done in the thesis explaining the details of the four summarization algorithms used.

Chapter 6: Gives the existing evaluation measures and the details explanation of the implemented evaluation method. It also gives the results achieved of the study conducted.

Finally we end the thesis work giving the conclusion done on the hypothesis of the thesis study and also gives a brief suggestion for future work in both text summarization and evaluation methods in Chapter 5.

1. BACKGROUND STUDY

In this section, we give the general history behind text summarization in terms of literature review, exploring previous work done in the text summarization field. This section also includes the fields that relate with text summarization and are used in the techniques for text summarization this are Text Mining, Natural Language Processing and Machine Learning.

1.1. Literature Review

Text summarization as a problem in or a field of natural language processing is defined as a way to get the subject matter of the source text without the need to go through the extra information provided. Text summarization was first implemented in the late 50's with the research dying down a little in the time between then until when the advent of internet and big data hit the world of research. In this section we will look at the first documented implementation of text summarization and the different ways it is implemented in recent years as well as the historical timeline of this field. The methods used in text summarization have evolved since the time it was first introduced by Luhn [2]. While sophisticated methods have been proposed over the years ranging in both complexity and quality results, some of the works have not veered off completely from the original application but only enhanced the ways to get the results. In this section we delve into the work that has been done in this field and where we currently stand in the text summarization problem.

1.1.1. The first documented research in text summarization

The first documentation of text summary was by the grandfather of text summarization, Luhn in the late 50's where 'abstracts' or what is termed as summary was extracted it was based on statistical information by measuring the frequency of appearance of a word in a document [2]. The frequency of the word determines the significance of a sentence, where depending on the words constituting a sentence it was given a significance score.

The sentence with the highest score is included in the summary sentences and the next sentence is picked accordingly until the desired summary length is reached.

As early as this work was done Luhn still saw the significance of giving machines the complete work of calculating the sentence summaries stating the human input is only required in preparation of the program. Luhn justified using frequency of word as a significant of importance by arguing that as a writer tries to give across the content of article, the repeated words form the basics of elaboration and the more a particular word is repeated in a certain way the better it relays the significance. If the words are repeated in more than one instant together then that signifies an important word. However common words that have otherwise no intellectual significance are excluded from being given significance priority, this words are now termed as 'stop words'. In this implementation no weight was given to semantic meanings of words, word with the same stem were treated as the same words e.g. go and going are treated as go same word. In the creation of this algorithm simplicity was the main driving factor Luhn stated:

> "The more complex the method, the more operation must machine perform and therefore more costly will be the process" [2].

Considering the time this algorithm was implemented this argument was highly sensible. The frequency method worked well because of the given nature of the field of application i.e. technical articles. Where Luhn stated synonyms don't make much difference because there are limited options and the author will switch back to the word after minimal attempt to change the wording.

With time other researchers did different approaches to text summarization, where single document summarization was the main focus of the text summarization field when it was a new research field, after the first introduction of Multidocument summarization research began to delve into text summarization using multiple text inputs[5]. Now most of the single document research is relegated to the more advanced abstractive summarization techniques as improvement is sought in the area while most extractive methods deals with multiple document summarization and the ordering of the sentences in the summaries. Also the type of input document switched from being

purely technically to inclusion of news summaries, short messages summaries (social media) and video summaries [6-8].

The languages also picked up given that in the onset of text summarization only English language documents were summarized and advancement were made, however as the research in this area picked up the momentum soon summarization in other languages soon picked up like in a new method for Arabic text summarization based on graph theory and semantic similarities between sentences was introduced, Maximal Marginal Relevance (MMR) method was used to do away with redundancy [9]. The authors stated that this was the first implementation of using a combination of graph method and MMR. In an enhancement of existing key phrase based method was able to produce a significant influence for Bangla text summarization over existing keyphrase-based methods [10]. Not only was there implementation and success in the summarization of other languages there was also multilingual summarization possibility where MEAD could deal with six different languages [11]. And in paper introduces a bilingual (English and Hindi) unsupervised automatic text summarization [12].

The methods also applied in text summarization also varied from the first frequency based summarizer, even though enhancement of this method are still being implemented, we have different methods like: use of events in summarization like the proposed method explored a frame-work improving phrase-guided centrality based summarization model that included a two stage summarization method, the first phase entailed the extraction of key-phrases and phase two using the key-phrases to get centrality as relevance model retrieval [13]. Three different methods of integrating events were proposed where it entailed the filtering of non-events, using event finger-print features and combination of the two methods stated above. Another method was by use of subevent for Multidocument summarization where it was proposed breaking down the documents into the subevents of the main event will help capture the sentences which are more relevant to the main event being discussed in the source documents [14]. Another method is use of Topic Signature, a set of related words with associated weights organized around head topic [15].

The topic signatures can be used in text summarization to get or identify complex concepts. The summarization is done by calculating score of sentences based on the relation to the topic signature the higher the relation the higher the score.

Also use of leading texts like in a paper that applied on news articles where a user is limited to queries in leading text to aid in better precision most news article produce a good summary using this method [16]. Machine Learning Trainable Algorithms have also grown in popularity where seeing that Machine learning has been applied in the different fields of natural language processing, it was bound to be implemented in text summarization as well, after it was first introduced and later proven that it could yield successful results, machine learning has been applied in text summarization by different methods [17-18]. In proposed a method where extraction of sentences was approached as a statistical classification problem where a classifier was trained to identify probability of a sentence being a summary sentences [17]. Another method employed a set of features which were extracted from the original document to train a classifier in identifying summary sentences implemented using naïve Bayes and decision tree algorithms [19]. In another method, approached the features as vectors and computed similarity between the features as well as values [20]. This paper uses word2vec to represent a word and neural networks model to generate each word of the summary, ONSES consist of three phases: the clustering phase which consists of clustering short text by means of the K-Mean algorithm then the second step which by using a graph based ranking algorithm rank the contents of the individual clusters and finally using neural machine translation in generating the main points giving successful results [7]. Use of lexical chains is also another method, Lexical chains are a collection or a sequence of related words. Summaries were picked in by choosing the concepts represented by the strongest lexical chains which gave a better indication of the central topic rather than when simply picking word in a text. Three alternative ways in choosing the sentences were given where it was either choosing the first sentence that the lexical chain are first spotted in, or choosing a representative word that on its first appearance the sentence it appeared in is chosen or finally choosing sentences with highest density of chain members which was the best option as the first two methods gave poor results [21].

Paragraph Extraction instead of sentences where the argument given arguing that that coherence will increase when paragraphs are picked over sentences [22]. Using Fuzzy Inference, the work in this integrated fuzzy logic with traditional statistical approached in a method that seemed to mimic the human mind when doing summarization [23].

Hybrid Summarizers have also become popular where their usage have successfully proven to be significantly efficient. A summarizer could be termed as a hybrid summarizer when it employs more than one method in summarizing a document, the summarizer could be a mix of extractive and abstractive or a mix of two extractive or abstractive techniques. A method employed a two phase extraction of summary from long text by first using a graph model in extracting the key sentences and generating a summary by feeding the extracted sentences into a recurrent neural network (RNN) based encoder-decoder model in order to get the model summary in [24]. While the authors approached the summarization as a two phase problem where they tackled keyword extraction using the successful TextRank algorithm and approach the second phase of getting sentence similarity using LexRank, this approach outperformed individual performance of the methods in [25]. In another work abstractive summaries were also produced from extractive method using WordNet ontology giving good results [26].

Text summarization has also borrowed from other field of natural language processing just like how it is normal to have interactions between the different fields where advancement in one field is applied on the other to see if the results are as successful example of this is use of Relevance Measure and Latent Semantic Analysis: Relevance measure is using basic information retrieval standards to get the relevance of a particular sentences in the document the highest most relevant sentence is extracted and the terms associated with it plus the retrieved sentence is deleted from original document after it's addition to the summary sentence, then the highest sentence is calculated again and the process repeated until the desired summary length is achieved.

LSA is also borrowed from information retrieval where depending on the singular value decomposition matrix the highest sentence in the highest concept is retrieved until desired length is achieved.

Taking from Sentiment Analysis: Sentiment analysis has found a lot of success as a field of natural language processing with the advent of social media hence the trial to see if it can also be successful in aiding other field. Sentiment of a sentence was used to see whether important sentences in the document could be detected in [27].

In this thesis also sentiment analysis was used to reduce the size of the input document arguing that if one sentiment outweighs the other no need to include the weaker sentiment sentences.

1.1.2. Important timelines in text summarization history

After Luhn first work on text summarization other works started coming up in text summarization some of the earliest work from late 50's until early 2000's included:

Edmundson text summarization where the method proposed entailed using a combination of features to get summary of technical documents, the feature sets included cue words, position and frequency of the word [28].

The first summarization done on commercial news where summarization was sentence level and depended on word frequencies [6].

The first trainable method was introduced where they used naïve Bayes to train a classifier in identifying sentences to be included in summary by [17].

Introduction of the first Multidocument news article (SUMMONS)summarizer in 1995, this was an NLP summarizer that summarized multiple articles that were based on the same event, the summarizer based on traditional language generation architecture had two main components, content planner i.e. the main process that identified what to include and what to discard based on a knowledge base and the linguistic components that selected the words that best described a concept and arranging the words such as to form coherent sentences [5].

In 1997 and 1999 proposed a second method for Multidocument summarization and graph based method for text summarization [29, 30].

In 1997 introduced scoring chain [31].

A salient based extractive summary in six languages was introduced which was a salience based extractive summary that is now popular in the field known as MEAD summarizer. The summarizer consisted of three components a feature extractor, a sentence scorer and a sentence rescorer by [11].

One of the application of mead on another summarization method is where they applied the mead summarizer in getting summaries of online news [31].

In 2001 introduced summarization using Hidden Markov Model a statistical tool used in modelling a generative sequence based on a previous observed sequence [32].

In 2001 also proposed LSA based summarization which was successfully used in the field of information retrieval, the application in text summarization proved also successful [33].

In 2002 introduced the Maxentrop i.e. the successful use of maximum entropy classifier in sentence extraction using with an optimised prior [18].

Overtime some of the most popular algorithms to be made readily available and used as base summaries are

TextRank**:** TextRank is an algorithm based on PageRank algorithm used by google, the algorithm is an unsupervised and based on weighted graph. The TextRank first pre-processes the text and then converts it to a graph with the weight of sentences i.e. similarity acting as the edges between them, after this step the PageRank algorithm is run through the graph and sentences with the highest weight are picked as summaries [34].

PyTeaser: A python implementation of TextTeaser, this is heuristic model which takes the linear combination of content selection features for extraction and chooses the sentences for summary accordingly [35].

LexRank: Also a graph-based unsupervised approach like TextRank, additionally does post processing step after extraction of summary to ensure minimum or no similarity in extracted sentences [36].

Latent Semantic Analysis: an implementation that uses singular value decomposition to get the ranks of concepts in the input text and extracts sentence accordingly [33].

1.2. Introduction to Text Mining

Text mining is defined by as a knowledge intensive process which a user interacts with a document collection over time by using a suite of analysing tools [37].

While described as Discovery and extraction of interesting, non-trivial knowledge from free or unstructured text in [42]. Another close definition of text mining is process of analysing text to extract information that is useful for a particular purpose where is text here is termed as unstructured amorphous and complicated to deal with [38]. Text mining is defined as Discovery of knowledge from database sources containing free text is called text mining in [39].

Using the term 'mining' with 'text' simply means that there is a hidden jewel that needs to be found. The underlying meaning of text mining therefore indicates that it entails getting the best information from the provided large source of text.

With the rise in technology in both hardware and software aspects, there has been an increase in data availability, web and the internet has changed the data that could be mined forcing data mining to evolve, where it was used to analysing structured data, now there is natural language flowing causing an increase in unstructured data, and hence the need for text mining techniques and algorithms which help the machine and the users to analyse and digest the information for decision making going beyond just information extraction. The reason why research has started regaining momentum in is credited to the power of text in the world of big data. Stated:

> "In these times the ability to extract information from many and disparate sources of data will help determine, in part, the balance of power between co-operations and among nations" [40].

Text mining uses techniques from data mining, machine learning, natural language processing information retrieval and knowledge management to deal with the information overload, it involves analysing of unstructured data by pre-processing the data i.e. turning into a data that can interact with machines then stores this data in an intermediate storage which after applying analysis techniques i.e. identifying the

pattern in text and determining which text to discard and which one to keep, helps visualize the results.

Text mining although a derivation of data mining, unlike data mining, deals with unstructured data and can be hard due to various reasons, some are discussed as follows:

Natural language is ambiguous in nature, a sentence could be taken to mean one thing while in essence it means completely the opposite, and there are different types of ambiguity in natural language which will be discussed in sections to follow. Natural language is also subtle and consists of misspelling and abbreviation especially with the advent of social media, world of emoji and abbreviated text. There are various dimensions to a certain word and concept and extraction leads to the questioning of all this dimensions. This are usually dealt with using Natural Language Processing techniques (details of this is discussed in NLP section)

Text mining tasks can be termed as either classification where a set of input documents are classified into their respective topic area what is termed as supervised learning because predefined categories exist. It could also be termed as clustering or unsupervised where documents that are analysed have to be grouped based on their underlying similarities and no prior existing categories but rather a learn as you go approach. It could also be in terms of information retrieval where meaning of multisource unstructured data is established.

1.2.1. Text mining use cases

Text mining has different use case which mostly can be categorized in terms of their target user base [41]:

For human consumption: The output of the retrieved information is for the sole purpose of humans rather than computer and hence no literal actionable results, this examples are that of text summarization, where a user bombarded with extra information and a piece of information required, this user may not know how to sift through the non-important stuff to get to the important information, text summarization therefore is used in getting the user brief snippets of important information that may

otherwise be a difficult or next to impossible to achieve. This is where text mining techniques are used to get consumable information. Another use case is of document retrieval where based on a user query a document containing the need the user requested is given as an output for the human, another is in information retrieval which is regarded an extension of document retrieval where instead of full documents a user is given only a snippet of the most important part of document meeting the needs of the user without bombarding with unnecessary information.

Assessing Document Similarity:

Text mining problems address this aspect when dealing with text by either using categorization/classification or clustering to group together similar documents. Or detecting the language of the document and grouping similar documents that have same language. Cases of also authorship identification come under this use case.

Extraction of Structured Information:

Entity extraction like name of people, dates, organization, and email addresses, etc. from unstructured data. Although they can be directly taken from data, having a large set of text it becomes difficult task. This can be tedious task in itself given for example in date retrieval, tomorrow , 2nd of January 2018, second day of the year could all be leading to the same date to a normal user but for a computer this is a difficult task.

Information extraction may be termed as getting the information needed to fill the blanks in an already existing templates like filling in details of the what, who, where of events in news articles. Steps in information extraction include fast getting the entities and then the relationship between those entities.

Learning rules from text:

This is one step addition to information extraction where the extracted information is used in learning to characterize the content of the text.

Text mining tries to solve the problem that is created by the existence of unstructured data:

"An important problem of mining textual information is that in this unstructured form it is not readily accessible to be used by computer" [42].

Structured versus unstructured data:

There is a huge difference between structured and unstructured data, while structured is an easily understandable data format unstructured is more of natural language which needs to be pre-processed before it can be used for any data mining purposes. Structured data are in a form of rows and columns which can be ordered by data mining algorithm. Unstructured data however is more like the human language and has expanded in use with the development in technology, social media texts or files in pdf and word format, audios, videos and images are some examples, this usually are hard to analyse due to the fact that human language is not straightforwardly understandable by a machine and ends up confusing it. Dilemmas often occur in categorizing data in terms of structured or unstructured data, data in forms of email could be categorized as structured from the way it is arranged with the to and from in place but because of the body of the text has an unstructured pattern due to the natural language presence.

Even though text mining is run through unstructured data, the text mining algorithms are not applied blindly, the bulk of text mining largely depends on text pre-processing techniques that prepare the data for manipulation done by complex or simple text mining algorithms. The figure shows the general architecture of text mining.

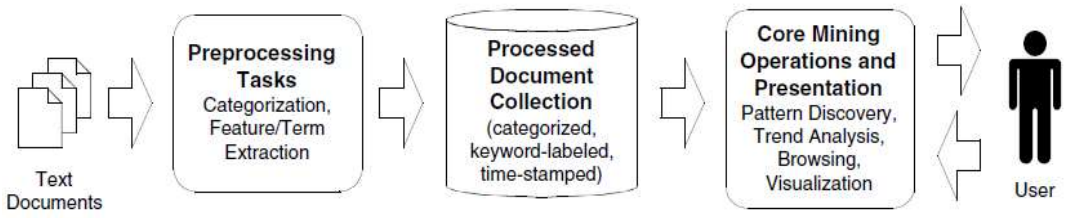


Figure 1.1. General architecture of text mining [37]

1.2.2. Difference between datamining and text mining

Data mining employs of approaches combining both artificial intelligence and database mining techniques to search through large dataset in order to get the underlying patterns and present it in a simple understandable structure.

The steps that datamining text in achieving this is [43]:

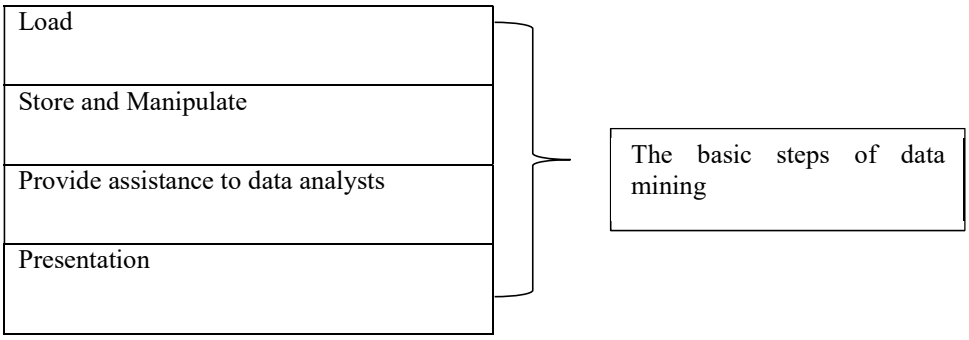


Figure 1.2. Basic steps of data mining

-collection, extraction, transformation and loading of data
-while in storage managing the data
-help organization or persons like data analysts access the readily stored data and depending on their needs help present the best way to organize the data.
-final step in datamining entails the visualization, the presentation of the output of the analysis done should be an easy to understand method like in tabular or graphical form.

Text Mining also known as Text Data Mining was before not given much consideration, traditional tools in data mining dealt with numerical data and the understanding that text also drives numbers in companies and the real world was yet to captured I terms of importance. When big data came about, customer feedback being at the tip of the hands and everyone giving opinions on matter, text mining started gaining popularity again in the field of research. Text mining techniques involve manipulating text using keyword, linguistic and statistical technological techniques to get breakthroughs in text analytics and that is by providing valuable structured information from unstructured text documents or resources.

Table 1.1. Differences between data mining and text mining

	Data Mining	Text Mining
Concept	Searches for underlying patterns and relationship among large stored data	Transforming and analysing of unstructured data to get meaningful structured data that can be manipulated and explored
Data Retrieval	Using standard data mining techniques to get the patterns in numerical data	Using standard text mining methods explored the lexical an syntactic features in text data
Type of data	Easily accessible homogenous structured data	Complex data structure which are heterogeneous and unstructured although sometimes may contain numbers, dates it's typically from forms such as articles, website, text, etc.

1.3.Natural Language Processing (NLP)

NLP is computer science field working in bridging the gap between human language and computer language, it overlaps between Artificial Intelligence, computer science and computational linguistics, that by use of machine learning algorithms, unlike hand-coding large sets of rules, generally focuses on interaction between human language and computers in an effort to increase machine's ability to understand or mimic understanding of human language [44]. It aims to reach a certain level of perfection whereby when a human is interacting with a machine, it should feel and seem so natural as if the machine was an actual human. NLP can be used to extract and analyse or perform tasks like text summarization, sentiments analysis, speech recognition and topic segmentation, by analysing the given texts it enables machines to interact with human in as close to humans as possible trying to bridge the gap between humans and machine in such a way that it will be highly unlikely to distinguish between the two.

It is described as the attempt to extract a fuller meaningful text representation from free text (human natural language) in [42]. It is employed in text mining by use of a set of linguistic concepts the likes of part of speech tagging (POS) and grammatical structure to turn unstructured text to structured text. NLP as a field started way back in the early 1950's with the first machine translation from Russian Language to English [45]. But the research was going too slowly and receiving negative reviews, the proposed research was not going as foreseen to the extent of limited funding for this field. Since that time NLP has grown as a field and is now among the top researched fields especially after the blooming of social media and the potential it gives having a rich content, in the past decade alone there are many publication in Natural language processing in fields like text summarization and sentiment analysis. NLP deals with different aspects of natural language like phonology, morphology, syntax, semantics and pragmatics [45]. It is working towards a point where the computers or artificial intelligent agents can pass the Turing test (a test of machine's ability to exhibit intelligent behaviour equivalent to or indistinguishable from that of human beings), main challenge currently being ambiguity of natural languages. This fields also shares similarities with HCI [46]. Despite the fact that they both aim to make interactions with computers more natural, there seems to be no much research conducted to check the intersection. NLP task is divided into natural language understanding and generation tasks, Natural language, in order to be transformed to a machine interacting structured language has fast to be pre-processed by use of the algorithms in natural language processing, task pre-processing is discussed below:

1.3.1. Text pre-processing

An integral part of NLP which deals with converting a raw text file into linguistically meaningful units. Natural language as stated, in its raw form provided lots of challenges when needed to be manipulated by machines, and with social media and growing technology most of the corpus are in natural language state in [47].

There are two stages involved in task pre-processing, the first stage is what is referred to as the document triage where it involves converting digital files into text documents that are well defined, this is where encoding conversion, language detection partition of text to allow discarding of unwanted part of the source text, this parts are like the links, headers and HTML formatting of the text [47].

The output from this stage is ready for analysis which leads to the second stage text segmentation, separating input text into word and sentence component which allow for individual analysis of the contribution of the concept in the whole document. There are two types of segmentation word segmentation or tokenization breaks the input text into singular words referred to as tokens based on boundaries between the words, this is followed by text normalization which puts the tokens into similar groups according to the base form of the word, an example are the word “MR”, “mr” “Mister” and “Mr” are all normalized into one group. Similarly the second type of segmentation, sentence segmentation, breaks the input document into individual by determining the sentences boundaries which are usually marked by the presence of punctuation marks. After cleaning most of the work done is shown below:

Sentence Tokenization: this is where the input document is divided into individual sentences for example the text:

“My son we have been meaning to tell you we are coming to visit. We want to see our grand kids and your new home. Expect us at the end of the month.” This simple three sentence text will be divide into individual sentences separated by the punctuations, the good thing is that in python there is a library called Natural Language Tool Kit (NLTK) which does this processes automatically when applied. We get the tokenized sentences as

“My son we have been meaning to tell you we are coming to visit”

“We want to see our grand kids and your new home”

“Expect us at the end of the month”

Word Tokenization: The text input can also be tokenize i.e. grouped into individual words for further pre-processing like stemming or lemmatization, in order to achieve this the text or sentences have to be broken into the main make up of words. NLTK has libraries (preconfigures programs) that perform the common tasks in NLP.

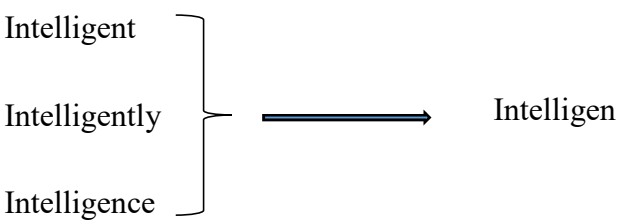
If we take our example of the three sentence input we gave above word tokenization will give: “my” “son” “we” “have” “been” “Meaning” “to” “tell” “you” “we” “are” “coming” “to” “visit” “.” “we” “want” “to” “see” “our” “grand” ”kid” ”and” ”your” ”new” ”home” ”.” ”expect” ”us” ”at” ”end” ”of” ”the” ”month” ”.”

Stemming: The three sentences below illustrate the example in what is done in stemming: Sentence 1: always work intelligently, Sentence 2: intelligence has always been admirable, Sentence3: she is always doing something intelligent

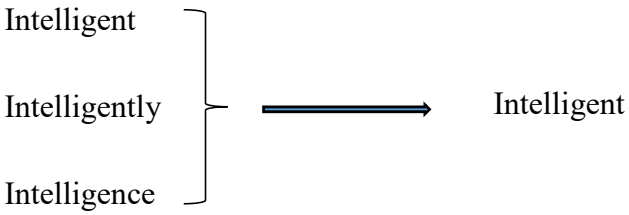
Sentences broken down to their word tokens:

Sentence 1:	sentence 2	sentence 3
Always	Intelligence	She
Work	Has	Is
Intelligently	Always	Always
	Been	Doing
	Admirable	Something
		Intelligent

If the word intelligence is taken from the above sentences we can have three forms of the word which are intelligent, intelligence and intelligently we get the stem form of the word as



Lemmatization: Lemmatization on the other hand will change the representation above to a more understandable method which given the lemma the meaning can be understood which can be shown as below:



Lemmatization and stemming both have their use cases, sometimes choosing lemmatization over stemming makes the work efficient if the meaning of the words are needed for the task at hand for example in a question and answering implementation. Stemming however has importance when there is no need for meaning of words like in the case of spam detection and filtering which we can do away with the computational effort that comes with lemmatization.

Creating a bag of words: Suppose our document has three sentences as Sentence 1: I love reading, Sentence 2: I hate Reading and Shopping, Sentence 3: Shopping is my hobby and my passion. Our Bag of words model is as below:

Table 1.2. Example of bag of words model

Sent vs terms	I	Love	Reading	Hate	And	Shopping	Is	My	Hobby	Passion
I love Reading	1	1	1	0	0	0	0	0	0	0
I hate Reading and Shopping	1	0	1	1	1	1	0	0	0	0
Shopping is my hobby and passion	0	0	0	0	1	1	1	2	1	1

Term Frequency Inverse Document Frequency (TF-IDF)

Bag of words gives us a simple representation of which words are present where however it does not give us any semantic information or the importance of certain words over others. To get a more meaningful representation we need calculation of the weight of word where some terms will have more importance than other, for example in the above bag of word the word ‘my’ has more weight than ‘love’ or ‘passion’ or ‘hobby’ when in fact this words should carry more importance therefore this leads to the introduction of TF-IDF model calculation. TF gets the local importance of a word if it appears frequently in the particular instant then it must be important [48].

Document frequency on the other hand determines the frequency of the term in the whole document, a term will lose its weight if it appears everywhere in the whole document indicating that the word is a common word and not just for the particular sentences in this instant it can be seen as penalization of common words while appraisal of common unique words.

From our example sentences above, we can calculate the TF-IDF by getting the tf from the bag of words model, the frequency of the term is termed as the TF.

IDF however is gotten from a calculation applied on the bag of words model. To get the significant terms in a document then we need to calculate the product of TF and IDF which gives us the values shown in the TF-IDF table example for the sentences in our example:

Table 1.3. TF-IDF Calculation example

Sent vs terms	I	Love	Reading	Hate	And	Shopping	Is	My	Hobby	Passion
I love Reading	0,18	0,48	0,18	0	0	0	0	0	0	0
I hate Reading and Shopping	0,18	0	0,18	0,48	0,18	0,18	0	0	0	0
Shopping is my hobby and passion	0	0	0	0	0,18	0,18	0,48	0,95	0,48	0,48

From the above calculation we can see that the word Love is given the highest importance in the first sentences while hate is the significant word in the sentence two and in sentence 3 the most significant word are my, is, hobby and passion which indeed express the most significant aspects of the sentences respectively.

1.3.2. Important terminologies in NLP

Phonology: use of sound in a particular language, this is the part of linguistic which refers to the system arrangements of sounds, a quick search for the definition of the term will give it as “the system of constructive relationships among the speech sounds that constitute the fundamental component of language i.e. study of phonological relationship within and between different languages”

Morphology: smallest units that form the different part of a word.

Lexical: this in relation to the words or vocabulary of a language.

Syntactic: uncovers the grammatical structure of a sentence and illustrates what a sentence tries to convey.

Semantic: Determines meanings of word or sentences in itself or in relation to other words or sentences in a document.

Discourse: Conveys meaning of components of sentences, not at a sentence or word level but at a document level. Helps in getting the main idea or clear picture of a document.

Pragmatic: reliance on world knowledge to get the goal of a certain word or sentences.

Information Extraction: obtains structured data from unstructured or semi structured data sources to get information that can be used by a computer.

Named Entity Recognition (NER): identifies and tags elements in a sentence according to a pre-defined set of categories like, geometrical location, event and people.

Corpus or Corpora: A large collection of textual documents that are used in analysis and mining of text data, this is the term given to the data that forms training and testing sets.

Sentiment Analysis: deals with identifying the opinion or sentiment value of text towards a particular subject in terms of positive negative or neutral sentiment.

Word Sense Disambiguation: Natural language is full of ambiguity (discussed in detail in the following section) that a computer cannot identify and would otherwise categorize wrongly, word sense disambiguation is a computerized way to do away with word level ambiguity, and usually this is done by use of knowledge base like WordNet or Wikipedia, a word like “tablet” is disambiguated in terms of that sentence to either stand for the gadget term or a form of a medicine (pill).

Bag of Words (BOW): this is a virtual bag created in text classification where individual words in a sentence are represented as multisets of words, using frequency of the word to calculate importance and train classifiers.

Latent Semantic Analysis: this is the analysis of relationship between a document and it’s terminologies with assumption that words that have similarity in meaning will also appear together in a text close together.

1.3.3. Challenges in NLP

The challenges in NLP stem from the fact that the field deals with natural language, therefore the problems in natural language are same for NLP, the challenges include:

Ambiguity: this is where a sentences or a word can have more than one meaning and this is a challenge because unlike the human mind, computers are not able to get the underlying meaning without a pre-set way of classifying it. Ambiguity in Natural Language can be at lexical, syntactic, semantic or pragmatic level.

Lexical ambiguity: A word can be ambiguous if it has more than one meaning in a sentence, lexical ambiguity is sometime easily solved by application of part of speech tagging where words in a sentence are categorized into their grammatical structure, but a challenge occurs where the category alone will not be efficient in ambiguity elimination, this problem is termed as lexical semantic ambiguity, consider the words bank, tank and pen, in this two sentences “The tank was full of water” and “ I saw a military tank” the word tank is tagged as a noun in both instances but meaning is different to both, this is where the computerized disambiguation is implemented i.e. the word sense disambiguation method.

Syntactic Ambiguity: this is also referred to as structural ambiguity and can be divided into two categories: Scope ambiguity when operators and quantifiers cannot be placed and each can take precedence over the other to give multiple meanings in a sentence for example consider the sentence “A woman without her man is nothing” the sentence depending on the precedence meaning could either mean that a woman is nothing without a man or a man is nothing without a woman. The other category is attachment ambiguity where a sentence constituents can be placed in more than one place in a parse tree, an example “The girl rode a horse in a red pyjamas” it is unclear to a computer with no real world knowledge whether the girl was in the red pyjamas or the horse.

Semantic Ambiguity: words in a sentence could take more than one construct for example consider the sentence “Alifah was content with her gift and Nihal was too.” In this sentences it is unclear whether Nihal was content with her own gift or that of Alifah.

Discourse Ambiguity: Need a shared knowledge to disambiguate.

Pragmatic Ambiguity: this is where a sentence does not provide the extra needed information that helps in completing the meaning it intended to pass. Consider " I love you too" does it mean I love you like you love me, or I love you as well as someone else or as well as like you and all the other scenario presented in such a sentence and without another set of information to get the information needed such ambiguity remains.

Language Variability: Natural language is very rich in terms of message passing, two sentence can have different structural and word construct but could mean the same thing for example "where is the nearest hospital "and "can you refer me to a nearby hospital" all lead to the same answer, easy for the human mind to comprehend but a task for the machine.

1.4.Machine Learning Understanding

Machine learning as a field has grown in its usage but before looking at some of the popular use cases, this section will first define machine learning and outline briefly why it is needed. There are several definitions of Machine learning tech target defines Machine Learning as:

> "A category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed" [49].

Before delving into machine learning what an algorithm is, it is defined as a sequence of instructions that should be carried out to transform the input to output [50]. Also defines machine learning as programming computers to optimize a performance criterion using example data or past experiences it states that machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample.

Machine learning is an attempt to equip machines with the ability to analyse information like the human brain. The human brain from an early onset learns information like colours, numbers, faces etc. to form a basis of opinion and differentiation of one thing from another. The same techniques were borrowed in Machine learning where given specific types of inputs in a certain field, machines

could statistically analyse the data and produce corresponding outputs and updates when deemed necessary. Machine learning much like data mining techniques studies underlying patterns from data structure to get meaningful aspects in the field or problem at hand. There is a lot of ways humans interact with products that are enabled by machine learning an example of such is language to language translation, recommendation of websites like amazon, email spam filtering, security protection from antiviruses even small appliance like facial recognition in open of phone by owner. This applications were all trained to get the correct categories and give important useful feedback. Machine learning algorithms were classified traditionally as either supervised or unsupervised but lately with the advent of big data a new classification has come which is termed as semi-supervised machine learning.

In a supervised machine learning a human, the data scientist, gives categorized training data, i.e. first trains the algorithm to identify categories by giving it marked data for example if we are training a machine to filter out negative sentences from positive sentences, the machine will be first fed with input that are labelled as either positive or negative and using this inputs the machine learns the terms that represent negative from positive. After the machine learns it can then be employed in real world by now getting unlabelled data and identifying the sentences that are negative or positive. In unsupervised machine learning which is also termed as neural networks the machines are not trained beforehand with labelled data but instead have to learn in a much more similar way than the human brain. Using deep learning techniques the machine checks and learns important features from input data the more it grows, in the world of big data and social media this type of machine learning is gaining importance as the day goes by, examples of this type of learning is the Facebook newsfeed where user gets a customized timeline depending on scrolling habits, where for example if a user likes reading a post from a particular friend and takes a pause every time they pass and see a post to either read or like a post from this post, when the user logs in they will see posts from the friends the habitually read appearing as the first posts in their feed, this is the machine learning algorithm storing and learning from the user as they scroll through the feed. The habit of the user are stored as probabilities where they get updated and change with the user habits. The semi-supervised approach is a method that aims to get the best of the two with reduced likeliness of making a mistake in

learning, with big data its highly unlikely to label all data that is needed to train the algorithms, therefore in fields where there is huge unlabelled data, advantages of both the traditional methods are taken as hybrid where a machine is fed the little labelled data and the huge unlabelled data to learn from.

Different animal analogies are used to explain the task that is machine learning, in one example that task of learning by machine is analogized by how rats treat food they encounter the first time where in the beginning they take a small bite and depending on their reaction to it they treat it as an okay or bad food and any future interaction with the food is treated depending on the similarity with the tried food in [51]. This was in the case of machine learning where if we look at the implementation of email spam filtering, examples in the spam figure. The machine basing on the learned experience of which messages were marked as sperm with the user mark the incoming similar messages as spam also, after a while depending on this only might reduce the efficiency therefore machines starts learning which word to associate with spam messages and encountering this particular terms in a message was a red flag.

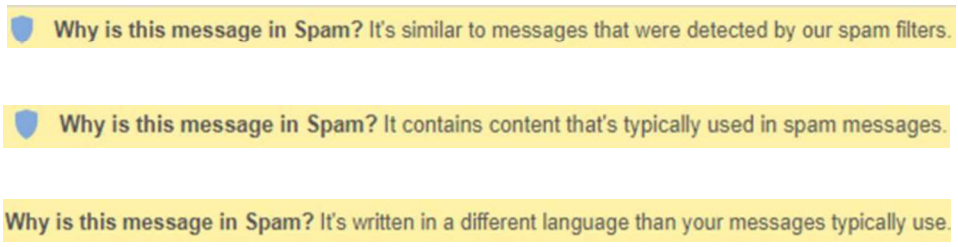


Figure 1.3. Gmail spam filtering examples

Machine learning can achieve effective results but it all lies in the training part, before training begins you have to choose which data to gather and decide which features of the data are important [52]. In order to have good performance from the machine learning algorithms the data used in the training should match the data that is supposed to be detected in the real world. If we take the human analogy in preparation for a biology exam, a human will not expect to perform well if instead they are reading text on history or another subject the same way if a machine is expected to for example get news article summary training using technical articles will not give the same results, that's why the same type is necessary but in doing so overfitting must be avoided where for example a machine is given a lot of training data such that it is hard to adopt where faced with unknown terms.

In defining how machine learning works it was stated that:

> "a computer program is said to learn from experience E with respect to some class of task T and performance measure P, if its performance in T as measured by P, improves with experience E" [53].

Going by this if we take a use case of machine learning based translation, the machine is said to learn from the interlanguage translation, the performance measure is by how accurate it gives the translation and the learning is determined if gaining from experience if the more it translates the better it gets as it interacts with both languages.

Over time, computers have learnt to do things like learning to recognize who gives what speech, which sentiment a user is giving, which books the people will look at next, which songs to recommend to a certain user, how to converse like a human and things like that.

It is stated that in order to have a well-defined problem, we must first identify these three features: the class of the task, the measure of performance to be improved and the source of experience [53]. This is true for any sort of learning that is done also in real life, a human for example has a task of living a normal life which is measured by how much improvement is done in his life and the experiences he gets from living are what help in navigating life. So taking an example of a machine language task of summarization:

The task: identifying which sentences are part of the summary and which sentences are to be discarded.

The measure of performance: similarity between the summary produced with the source article or any other evaluation metrics.

The source of experience: the identification process and learning of new features each time a new document is introduced.

Reasons to pick Machine learning algorithms from computer programs:

> "Machine learning is picked over normal programming when two aspects of problems are involved that is if a problem is too complex or there is a need to constantly adopt to changes" [53].

Complex tasks are categorized into two tasks, that which a human can do naturally but a computer can't and that which is too robust for even human beings. The task that a human can do but computers can't include tasks like speech recognition, i.e. identifying who gave speech and where and analysing the tones, identification of different faces or different animal species. The second type of complexity is where as a result of big data, there is too much information and little analytical skills and times needed to get the best advantage of existent knowledge hence machine learning need where medical histories or world happens can be analysed with a click of a mouse.

Flexibility of machine learning is needed where unlike the rigid computer programs that are written by a human to just execute a simple task, machine learning could learn and adopt to changes going far and beyond what it was taught in the beginning and the more it interacts with the data the more robust and accurate the output gets.

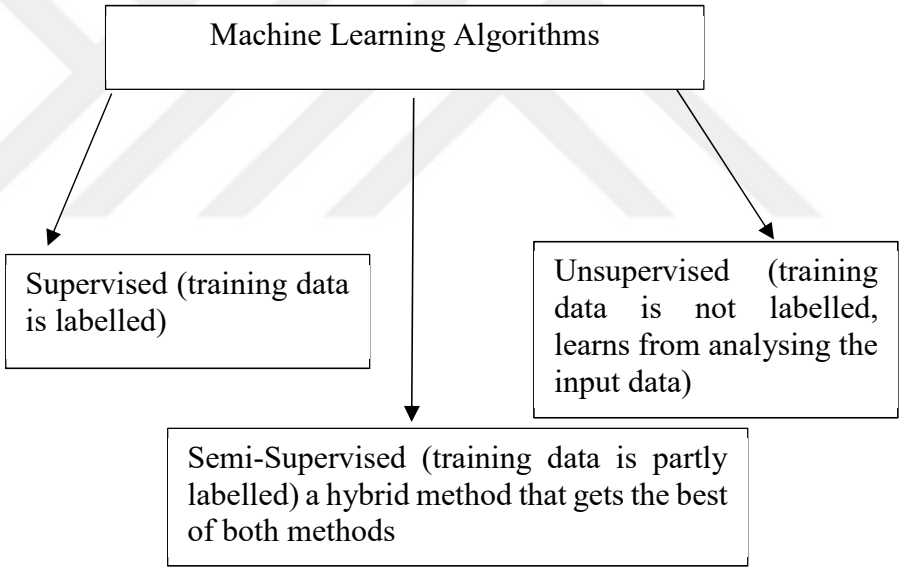


Figure 1.4. Machine Learning Algorithm types

2. TEXT SUMMARIZATION

A summary is defined in Cambridge Dictionary as a short clear description that gives the main facts or ideas about something [54]. While in oxford dictionary it is defined as a noun means a brief statement or account of the main points of something [55]. Macmillan Dictionary describes summary with terms not as far from the two a short account of something that gives only the most important information and not all the details [56]. In techopedia a summary is a process by which a computer program creates a shortened version of text [57]. In all definition it is noted that text summary should be, short, precise and give the general idea without the need to refer back to the original document. An ideal automatic generated summary will be one where the summary reads like one generated by human without the emotional bias. The summary in this scenario will eliminate redundancy i.e. no two similar sentences in a summary, have only important sentences and discard any sentence that doesn't directly give new information or doesn't elaborated on already given information and be coherent in that the selected sentence have to have a flow that the reader would not struggle to put together.

Text Summarization is a field in text mining that deals with reducing the information that is existing in source documents to a smaller perusable size summary that is both informative and quick while keeping the main idea behind source documents. Automatic summarization was introduces in the late 50's by Luhn in his paper that intended to give available online summaries of existing scientific document [2]. Over time it moved from scientific documents to news and blogs as technology grew. It's importance is seen especially in the current world where information is existent in billions but people remain less knowledgeable, a user has so little time to capture all those information and to decipher which one to keep and which one to discard, with the help of summarized contents, a user could automatically classify important information from unimportant according to his needs.

Technology has changed so many aspects of the human life that automatic summarization can be fully appreciated especially if the system outputs a summary

that would be self-sufficient on its own without need for reference document check. In the current summarization systems, automatic summary is a three-step process the first step being the establishment of the important sentence that need to be included in the summary and those that should be discarded, the ways the content to keep and discard are chosen will be discussed in later stage, while the second step is defining the way the chosen sentences will be ordered in the output summary while the third step is in the generation of the summary based on the chosen sentences, which entails cleaning up the sentences or leaving them as is or even rephrasing of the selected contents. While the former two are thoroughly researched and evolved area the third phase is still a work in progress with minimal achievement due to the difficulty involved and in part negligence. Text summarization in its basic form can be categorized into Abstractive, Extractive and Hybrid summarization, and each of this type can be further detailed in terms of targeted audience, output and input source, in the next part this summarization types will be briefly explained and in the next step abstractive and extractive text summarization will be discussed in details.

2.1. Types of Summaries

Summaries can be placed into types according to the:

Number of Input Document:

In the early summaries only single documents were used as the input source and summaries were extracted in terms of an abstract or an outline of the document or just a one sentence summary classified as the headline of the document, but introduced the first multi-document summarizer where summaries were now extracted from a news series of multiple documents covering the same event or blogs and webpages discussing the same topic or answering the same question [5]. Hence a summary could be categorized as a single or multi-document summarizer according to its limitation in the number of document inputs it can accept.

Number of Input Language:

A summary could either be classified as monolingual or multilingual based on its limitation to acceptance of a language, earlier work in text summarization were all based and limited to one language mostly the English text and later the summarization

was adopted to other languages like in Bangali and Arabic and with time a summary could pick source document from different languages and summarize it in the specified language summary. [10, 9]

Genre / Input Text Limitation

Summaries of a document could be defined as either constrained to a specific genre like scientific articles or news only or specific fields like biomedical or could be open template based summary of any field with no limitation of the domain of the input text [58,59]

The Target Audience:

Summaries don't have a specific point to address but a rather a general summary of the documents for any interested party with no targeted audience are referred to as generic summaries while on the other hand query-focused systems produce summaries of documents with respect to a user query and without the need to refer back to the original document(s) could surface as a question answering complex system.

Output Information

In this type of summaries, the output could be just an outline of the documents and a peak preview of what a document contains, giving the audience what to expect in a document and make a decision to either dig deeper and read the source document or discard them altogether, this is termed as an indicative summary, on the other hand there is the informative summary where the summary on its own could act as a standalone informative piece that doesn't require the user to dig deeper into the source document since it can provide all the important information and ideas expressed in the source document(s).

Output Generation Approach:

This categorizes summaries as either abstractive, extractive or hybrid summaries, extractive summaries have been researched more and are consistently used due to the fact that they are easier of the two to implement, this type of summary, chooses the sentences to use from a list if input sentences and discards the sentences it terms as

weak, the chosen sentences are then taken as is without changing any structure and extracted as the output summary. Abstractive summary on the other hand employ sentence generation techniques and the summary is given almost as close to human, the important sentence after extraction are sometime reconstructed or paraphrased to give the intended meaning or a more coherent sentence than the extractive summary would have. In Hybrid summarization, the system maximizes on abstractive summarization by use of extractive methods, in the extraction phase extractive techniques are used and the output is fed to an abstract generator which smoothens out the sentences rephrases and produces abstracts. The three types of summary employ all the other types of summary categories as per need and hence chosen in this work as the mother summaries in terms of summary categorization.

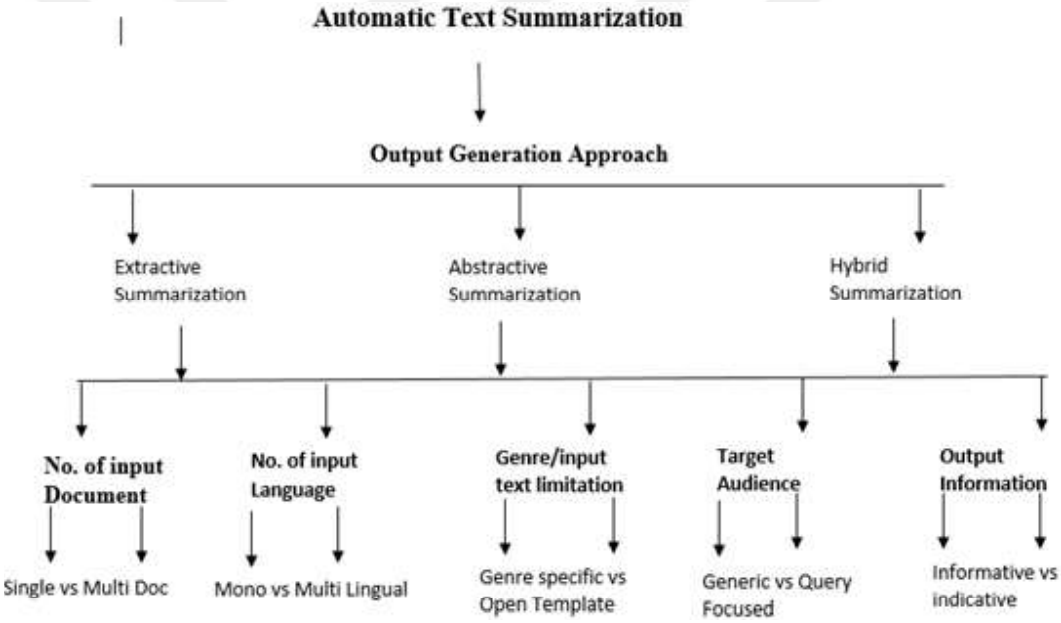


Figure 2.1. Pictorial view of summarization categories

2.2. Application of Text Summarization in Real World

When documents are summarized into a reduced version that doesn't lose the initial idea of a document, time needed to go through the billions of text documents reduces, this directly translates in the real world as of benefit to researchers and students working in certain area could get a feel of a document before time and effort is wasted in finding good sources for their work, in the same instance a journalist could effortlessly keep up on the latest in an event or news coverage around the world, an assistant could get discussed issues in form of bullets and in the health sector a doctor

could get the history of a patient from a summary instead of going through years of documented health details. In the works that have been done previously it shows that automatic summarization could effectively help in teaching for example, English teachers, in summarizing, extraction and analysis of information in quicker method, Google also benefits from text summarization in their search engines by keyphrase extraction. Other field of natural language processing also benefit from text summarization as in sentiment analysis instead of analysing huge corpus they instead reduce the size by summarizing and then doing the analysis in the resultant summary[60].

2.3. The Main Methods of Automatic Summarization

Text summarization methods are categorized into three which could either be an implementation of any of the above summary types. The categories are abstractive, extractive and hybrid techniques. The abstractive and extractive are discussed in details while the hybrid is mentioned in brief.

2.3.1. Abstractive text summarization

This part will focus on abstractive summaries. Their inceptions and advantages over extractive summaries.

In brief work done in abstractive text summarization can be said to be:

> "Abstractive summarization performs summarization by understanding the original text with the help of linguistic methods to understand and examine the text" [61].

Abstractive text summaries seek to eliminate problems that occur in extractive summary and give a close to human like summaries, the major steps involved are in the selection and retrieval of the content to use by use of basic features and reducing and rephrasing the selected sentence to extract a coherent summary that may not be constructed with the same words or phrases in the original text. When dealing with abstractive text summarization there are some major cut and paste work done in creating the summaries [62]. This work can be categorized as either sentence compression where sentences are reduced in size only keeping the important concept and discarding the extra details in either creation of shorter summary or even headline

generation. It can also be sentence fusion where sentences with similar content but different in some wording are fused to reduce redundancy but keep the important information from all the different sentences. It also could be reorganization of syntax which create coherence where otherwise after paraphrasing the sentences will be grammatically inconsistent. The other major work entails lexical paraphrasing where instead of complex terminologies a simple easier to understand term is used in place of a complex terminology. The challenge though with an abstractive summarization is what is termed as "representation problem" i.e. the generating system capability is limited to the richness of their representation and ability to generate such structures- a system cannot generate what their representation cannot capture. For a fully efficient abstractive summarizer there is a need for a system that can analyse and capture natural language with all its underlying meaning. Abstractive summaries are categorizes based on the approaches used to generate the abstracts, Structured based approaches uses cognitive schemas like frames, scripts and templates to encode relevant information from source document. Semantic based approaches information is taken from source document and restructured into a semantic representation which is the fed to natural language generator to create abstracts.

Structured Based Approaches

Rule Based Methods:

Documents are first classified into their respective categories and questions which form the basis for rule generation are asked accordingly, the rules are then used to place similar verbs and nouns enabling pattern generation by context selection, this patterns are then use to create the summary sentences [63].

Ontology Methods:

Also termed as knowledge based is a tedious process due to need for domain expert who define ontology for news events used in processing phase to produce meaningful corpus while discarding the rest.

Tree-Based Method:

This method applies the use of trees in that the original document i.e. the source text is represented in tree form which is traversed to get the central content and the sentence are selected based on the content, clustering algorithms are used here in forming the sentences after getting the centroid concept.

Template Based Method:

Templates are formed as a guide for the main extraction process. Based on the user requests the templates are filled with snippets and given as summary. However this is also a tedious process because it needs the manual creation of templates and in its own creation of all future possibilities of scenarios in terms of templates is not possible.

Lead and Body Phase Methods:

A method that relies on rewriting the lead and body sentences by substituting and inserting semantically related chunks from the original text.

Graph Based Methods:

Many researchers use a graph data structure (called opiniosis-graph) to represent language text. In graph based methods, the nodes and edges are used in representation of the structural words and connecting among them via weighted or directed edges.

Semantic Based Approaches:

Multimodel Semantic Models: captures sentences and forms relations between them expressed as sentences, it has three phases, the first phase is the semantic model construction, the model is created by using sentences as nodes and relations between them as edges or links where in phase two the density matrix is used to extract the most important concepts leading to phase sentence generation which in abstractive summary entails cleaning up the sentences or even rephrasing it mimicking the human summary generation.

Semantic Graph Based Models: a rich semantic graph (RSG) is created which is used for summary generation in three phases: the source document is fed into an RSG and

the verbs and nouns form the bases for nodes while the relation between them are the edges, in the second phase, by use of heuristic rules the graph is reduced to a lesser graph and in final phase abstracts are generated, this method produces more accurate and unique summaries i.e. less redundant and coherent it's limited to only single document summarization.

2.3.2. Extractive text summarization

This section will delve deeply into extractive summaries and explain how it operates.

Extractive summaries identify the most important sentences from the source and put them together to form the summary, it involves three steps which do rely on each other but work as an extension of each other. The first step involves the creation of an intermediate expression of the input text capturing only the important aspects, then the second step established a sentence scoring mechanism and gives the candidate sentences and final steps selects the candidate sentences at the top of the requirement based on the specified features and the sentence scores from step two and creates the summary consisting of sentences up to the length of the required sentence. Various approaches are employed in extractive summarization and this approaches are:

Term Frequency-Inverse Document Frequency:

The creation of a TF-IDF model was already discussed in the previous chapter discussing pre-processing methods. Summarization using this method usually scores sentences based on the weight of the term it contains, the weight of the term is taken from the tf-idf model created. The total of the sentence is the sum of the weight of the individual term it contains, the summary is taken from the sentences with the highest score each time picking the highest sentence after eliminating the already picked sentence. Often very long sentences are penalized in order to pick sentences based on content not based on the bias created by length.

Classical Method: this is a basic approach to text summarization where sentences are given scores according to four factors high content of frequently appearing word i.e. keyword content, presence of cue word, this are words that indicate a sentence is an expression of an important idea, close relation of the sentence to the title and heading

words and finally sentence location, the sentence score determines inclusion or discarding of the sentence.

Cluster Based Method: a cluster is a group of sentence that have similarities, sentences are grouped into this clusters and when a sentence is picked from a cluster it is compared to other sentence in the cluster to rate inclusion of those sentences.

Graph Theoretic: this method provides a theme identification possibility, sentences are represented in a graph form the node representing the sentences and the edges the similarity between the sentences, and the higher the presence of an edge in a node the more relevant the sentence is and based on this sentences with the highest edges points are selected for summary.

Rhetorical Structure Theory (RST): A tree based approach where relations among sentences are explored by use of tree representation, sentences form the node and are connected through RST relations. Important sentences are retrieved by tree-traversal methods and the top n sentences are extracted as summaries.

Machine Learning Techniques: By use of statistical measures fed to a machine, sentences are classified using machine learning algorithms as either summary or non-summary.

Challenges in Extractive Summarization:

Extractive summary does not check for underlying meaning in sentence but relies on features that term a sentence as either important or unimportant, one of the challenges occurs when long sentences are included due to their frequent word contents, this leads to unimportant parts of a document forming part of a summary and causing an information overload.

Important details might be missed out on because extractive summary will pick the top sentences and in long documents information is spread out.

Inaccuracy or misrepresentation of conflicting info when the nouns and proper nouns are mixed this is termed as dangling pronouns, when the missing piece is missing the information might be misinterpreted.

Extractive summary mostly luck a coherent structure because sentences are picked randomly and fixed together in the summary with no post processing phrases.

2.3.3. Hybrid text summarization

This type of summarization usually merges the extractive summarization simplicity of sentence extraction with the complex abstractive summary generation that produces a paraphrased version of summary. An example is where the proposed model in this paper is a two-phase approach towards long text summarization EA-LTS [24]. It consists of:

Extraction Phase: Conceives a hybrid sentence similarity measure by combining sentence vector and Levenshtein distance and integrates it into graph model to extract key sentences.

Abstraction Phase: It constructs a recurrent neural network based encoder and decoder and devices pointers and attention mechanism to generate summaries. Test is done on a real-life long text corpus and results verify the accuracy and validity of proposed method.

And also another example is where proposed a system that generates abstractive summary from extractive using WordNet ontology. The results indicated that the summarization were correct grammatically and in terms of readability [26].

2.4. Approaches to Content Selection

The basic and most important step in summarization is in choosing what to keep and what to discard, in what we term as content selection. It has been shown that the choice of the features chosen to extract the content depends on the context of the source document, mining data from different datasets can be maximised by use of the correct set of features [64]. Using three different context i.e. news, blogs and articles the paper evaluated techniques advocating that the quality of summary obtained using various techniques all depend on text subject where depending on the text one technique is more effective than others. The summarization method used was easy combination of different sentence scoring methods in order to obtain the best summary depending on the context. They evaluated using 15 of the most popular sentence scoring methods,

the combinations that will yield the best summary giving a score between 0 and 1. The proposed combination methods were in terms of:

A) Ranking: Every service selects the main sentence and the user combines it

B) By punctuation: The service scores each sentence and then returns one sentence with updated score.

The corpus used was three different set; the CNN data set for the news, blog summarization dataset and SUMMAC dataset for the articles. The evaluation technique was quantitative assessment Recall-Oriented Understudy for Gisting Evaluation (ROUGE).

The results showed different techniques work differently for three types of dataset, since the CNN dataset were formal and structured the one that worked was combination of best word based and best sentence based algorithms, i.e. a combination of word frequency,

tf/idf, sentence position, and resemblance to the title, the blogs were less informal and unstructured, they achieve good results using word frequency and tf/idf but in comparison to the news combining the methods with text rank score and sentence length gives an improved results, for the scientific were well structured too and the best combinations include: cue-phase, sentence position, tf/idf, and resemblance to the title.

Generally, in the selection of the content there are two types involved, unsupervised and supervised content selection:

Unsupervised: This approach dates back to the first works in summarization where sentences were chosen if they included informative/salient words, salient words are defined by two methods, TF-IDF and Topic Signature, Frequency based approaches and Feature Based Approaches [2].

Frequency Based Approaches for Content Selection:

This approach takes advantage of the assumption that a word that is repeated mostly in a document is an important word and hence its usage in a sentence will indicate that

the particular sentence is an important context to be included in a summary, the more the number of frequent word used in a sentence the highly likely it is to be flagged as important. There are two techniques that use this approach:

Word Probability:

Summarization is done on lengthy document hence just counting the number of times a word occurs is tedious and inefficient because it influences the occurrence of words, to cub this word probability is computed to give better impact in reliance on this feature. Word probability is computed and then use as an indicative measure for sentence weight.

Term Frequency –Inverse Document Frequency (TF-IDF)

Calculates the weight of frequently occurring terms in comparison to its frequency in the document.

Feature Based Approaches:

A sentence is ranked as an important sentence based on its score on the total number of features it contains, this features are selected from a category of features and each algorithm has a combination of different sets of features for sentence scoring, this include:

Inclusion of title word in a sentence: the more similar a sentence is to the headline of the document the more relevant it is to the summary extract.

Position of a sentence in a document: sentences occurring in the beginning of an article is more favoured than those in the middle, because mostly in the writing of articles the main ideas are first exploited and the middle is usually a drift from the main idea and included related items which are important generally but are often not relevant for summary consideration, the end sentences are sometime also important because in conclusion the main idea of the document is usually given. Sometime this feature fails because in some articles the first sentences are usually just introduction and a way to pull the reader into continuing to the important information in the body, but this is less in terms of occurrence.

Sentence Length: Longer sentences or very short sentences are usually discarded in favour of medium sentences with the assumption that better medium sentences hold weight and direct to the point while long sentences are an explanation of the medium length sentences, while the short sentences are bridges or what is termed as connecting sentences which don't give the main idea behind the topic in discussion.

Term weight: The weight of words in a sentence are calculated and sentences that have higher weighted terms are favoured for extraction.

Proper noun: Inclusion of a proper noun in a sentence is an indication that it contains contents of relevance to a summary compared to sentence that are plain.

This are some of the combination of the features are used in calculation of relevance of a sentence.

Machine Learning Approaches:

Supervised Approach: this is achieved by use of machine learning algorithms, first binary classifiers are trained and tested to identify features that classify a sentence as either inclusive or outside summary target, features like sentence position, cue phrases etc. And then later used to extract content to be used in summary.

3. COMPARATIVE STUDY FRAMEWORK

The text summarization in the thesis was done to test the arguments that indeed the summaries created by summarizers help the user in saving time and in this generation where users are bombarded with a lot of different articles and all kinds of information. And the second argument that indeed users can be used to evaluate summarizers. The study was done using four implementation of algorithms two already in existence and the two others an enhancement of existing methods. The summarizers were termed as Frequency-Based, Gensim, Sumy LSA-Based and Sentiment analysis-Based. They were applied on news articles taking from online sources. The summarizers all followed the following method:

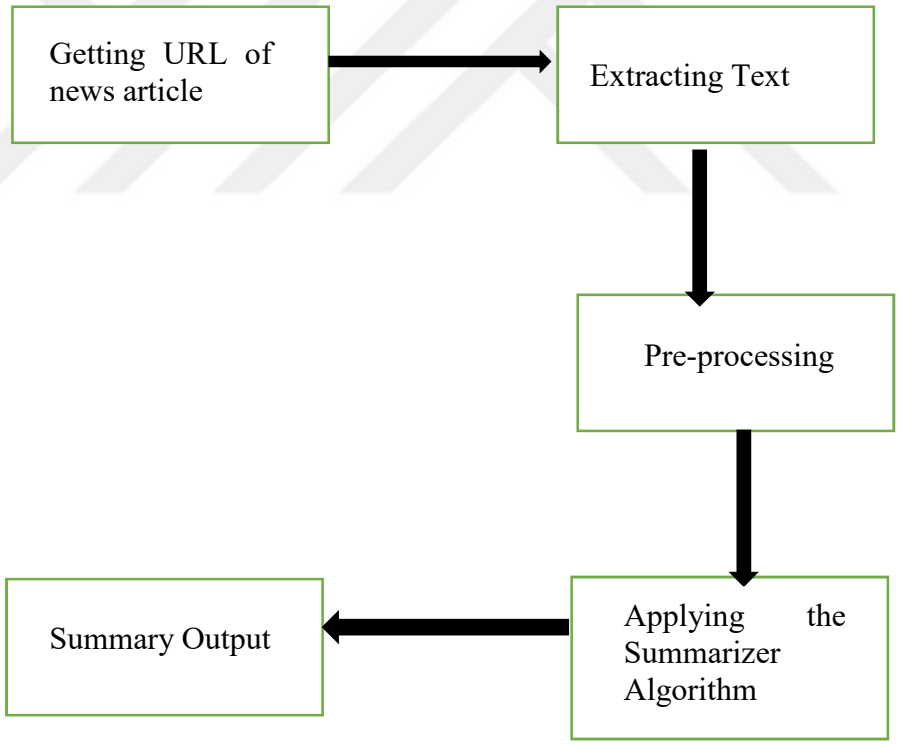


Figure 3.1. General flow of the summarizers applied on a news article

The four summarizers were implemented using Python programming language taking advantage of the rich NLTK Libraries that simplify the interaction with natural language. They are discussed below and the results of the work done is discussed in the result section.

3.1. Frequency-Based Summarizer

In this summarizer, the approach implemented is a statistical frequency based algorithm based on the original summarization techniques that applied frequencies to score sentences. The method approached a combination of the frequency method with simple selection of words that represent an occurrence of an event in order to enhance the improvement of the frequency based summarizers. The steps involved in the algorithm are stated below:

-Applying POS tags to the words in the input text and selecting only words with the tags NN, NNP, NNS, NNPS and VBG. This were termed ‘Allowed Words’

-Removal of Stop Words by use of NTLK library that provides an English list of stop words.

-selecting only allowed words which are not part of the stop words as the words that create the Bag of Words.

-Creating a Histogram from the bag of word

-Sentence scoring by adding up the scores of individual terms in the sentences.

-selecting summary sentences.

The idea behind this summarizer was set to pick out the top n sentences that best represented the articles and could easily help in the identification of the event in that particular article. To understand the scope behind events and event extraction we briefly discuss in the following sub-part what events are all about

Event Extraction

This deals with the extraction of the who’s the when’s and how’s of an occurrence. Which could be either past, present or continuous.

In the English Oxford Dictionary the term Event is defined as a thing that happens or takes place especially one of importance retrieved from the same website as [55].

Event Extraction Methods:

Event extraction is an important aspect in general information extraction, it entails a combination of different domains like computer science, linguistic, data mining, artificial intelligence and knowledge modelling. In the beginning event extraction was introduced as a way to monitor events revolving around terrorism, later on it spread to other domains with the likes of finance, politics and election benefiting from extraction of events. Currently with the growth in use of unstructured data, event extraction is useful in information extraction applications like risk analysis, decision making support tools and most recently in text summarization. Depending on the field of modelling, event extraction techniques can be divided into either Data-Driven, Expert-Knowledge Driven or a combination of the two techniques a hybrid extraction technique [65].

Data-Driven Extraction Technique: using statistics, data mining and machine learning techniques, input data is converted into useful knowledge.

Expert Knowledge Driven: by exploiting real world existing expert knowledge bases, mostly the pattern based approaches, events and information are obtained.

Hybrid: This achieves better results by combining both techniques and maximizing on knowledge driven techniques using machine learning.

This are all complex methods used in extraction of events but in this thesis, a method extracted events in a unsupervised method aimed at improving object and event monitoring, the paper pointed out that object and events are most likely to be a noun and gerunds (a verb function as a noun ending with an ‘ing’) it borrowed from [66]. After POS tagging, the words selected were those with the NN-Noun, NNP-Proper Noun Singular and VBG- verb, gerund or present participles. This was implemented in the frequency based summarizer in an attempt to give higher score to the sentences containing words with such tags.

This summarizer was then applied to a news article taken from an online source. The results of the summarizer are discussed in the Evaluation and Results section.

A random article was taken from the internet and the summarizer was used in providing the summary the result shown below give part of the article and the summary produced.

(CNN) -- Robin Williams -- who first made America laugh and eventually touched "every element of the human spirit" in a remarkable range of performances -- died at his Northern California home Monday. Williams apparently took his own life, law enforcement officials said. He was 63."He has been battling severe depression of late," his media representative Mara Buxbaum told CNN. "This is a tragic and sudden loss. The family respectfully asks for their privacy as they grieve during this very difficult time. "Coroner investigators suspect "the death to be a suicide due to asphyxia," according to a statement from the Marin County, California, Sheriff's Office. Williams married graphic designer Susan Schneider in Napa Valley, California, ceremony in October 2011 This morning, I lost my husband and my best friend, while the world lost one of its most beloved artists and beautiful human beings. I am utterly heartbroken," Schneider said. … The full article can be read from the source **[67]**

5 sentence Summary Produced by simple summarizer:

"An investigation into the cause, manner and circumstances of the death is currently underway by the Investigations and Coroner Divisions of the Sheriff's Office," the sheriff's statement said.

"Word of Williams' death stunned the entertainment community and beyond Monday. President Barack Obama's statement sent from the White House summed it up: "Robin Williams was an airman, a doctor, a genie, a nanny, a president, a professor, a bangarang Peter Pan, and everything in between.

"Coroner investigators suspect "the death to be a suicide due to asphyxia," according to a statement from the Marin County, California, Sheriff's Office. Williams married graphic designer Susan Schneider in Napa Valley, California, ceremony in October 2011.

(CNN) -- Robin Williams -- who first made America laugh and eventually touched "every element of the human spirit" in a remarkable range of performances -- died at his Northern California home Monday. Williams apparently took his own life, law enforcement officials said.

""Happy Days" star Henry Winkler said it was "unimaginable that this is the reality today, that this incredible human being, incredible, delicate, funny, dramatic human being is gone.

Figure 3.2. Results of the simple summarizer applied on a news article

3.2. Gensim Summarizer

The second summarizer applied is an existing application of TextRank as a library in python, the Gensim Summarizer implemented by using the Gensim library [68] provided in python which has a list of applications of summarization to choose from. The text to be summarized is given as an input as a text or URL then the required summarizer is called and implemented. In this thesis study the implemented summarizer from Gensim is TextRank summarizer.

TextRank was originally implemented and the Gensim TextRank summarizer is an implementation of this technique [34]. In TextRank a graph is constructed in order to find most relevant sentences in text. In the graph the edges represent sentences in a document and the connecting edges represent the relation between this sentences which is based in the content overlap which basically means number of common words found. The TextRank algorithm is borrowed from the Google PageRank algorithm that is used to rank web pages, this algorithm used the notion of graphs to get calculation for most important webpages. A high rank is given to a sentence with a link from a higher ranking page. If a sentence contains words that appear in many other sentences it is assumed to be of importance too. The TextRank implements a basic voting system with the sentences with the highest important votes are chosen as sentence summary. The voting systems depends on a) the number of votes a vertex gets and b)the importance of the voting vertex, if the vertex casting the vote has higher importance then it is a given that the integrity of that vote is higher. The steps in the original TextRank are as below [bringing order]:

1. Identifying the vertices of the graph, this means getting the textual pieces or units that best express the input text, it is in text summarization sentences.
2. Getting the relations among the chosen pieces in this instance sentences which then aid in the construction of the connecting edges which can either be weighted or unweighted, directed or undirected depending on the choice.
3. Iteration of the graph-based ranking algorithm until convergence.
4. Ranking the vertices based on the given individual score.

This algorithm implemented in Gensim was taken as is and used in summarization of news article as part of the comparative study of text summarization algorithm.

NAIROBI, Kenya — President Uhuru Kenyatta was declared on Monday the winner of Kenya's presidential election — for the second time this year. Mr. Kenyatta received nearly 7.5 million votes in the repeated vote, held last week, the national elections commission announced. Mr. Kenyatta also won the first election, in August, by 1.4 million votes. But the opposition leader, Raila Odinga, challenged the results, and the Supreme Court nullified the election in September, citing irregularities. Backers of Mr. Kenyatta interpreted both of his wins as broad national support for the president, but opposition supporters said they had twice been disenfranchised by a process that lacked credibility. Mr. Odinga withdrew from the second election two weeks before the vote, arguing that the electoral commission could not oversee a free and fair process, and he called on his supporters to boycott. His name nevertheless appeared on the ballot, and he collected just over 73,000 votes, compared with nearly seven million in August. Elections officials also cast doubt on the credibility of the process in the days before the vote. One commissioner fled the country and resigned, citing death threats and questioning the impartiality of the commission. ....full article from source [69]

Summary produced 20% of the article:

But the opposition leader, Raila Odinga, challenged the results, and the Supreme Court nullified the election in September, citing irregularities. Backers of Mr. Kenyatta interpreted both of his wins as broad national support for the president, but opposition supporters said they had twice been disenfranchised by a process that lacked credibility. Mr. Odinga withdrew from the second election two weeks before the vote, arguing that the electoral commission could not oversee a free and fair process, and he called on his supporters to boycott.

The top elections official, Wafula Chebukati, warned a week before the polls opened that political interference in the commission's work was likely to undermine the credibility and neutrality of the vote. Mr. Chebukati backtracked on that criticism while announcing the results on Monday, declaring the process "free and fair." At least 14 people have been killed in election-related violence since the Oct. 26 vote, according to international officials, and more have been injured.

"The killing was well planned and executed," he said, and was intended as "a direct warning to others." Mr. Odinga condemned the violence, which he described as perpetrated by Kenyatta supporters, but stopped short of urging his own supporters not to react with violence. Kenyans, foreign diplomats and even some of Mr. Kenyatta's own supporters have called on the president to engage with opposition leaders after his win, but in his victory speech the president distanced himself from dialogue. "Those who want to ask me, 'Are you going to engage in dialogue with so and so and so and so?' let them first and foremost exhaust the constitutionally laid out processes," Mr. Kenyatta said.

Figure 3.3. Results of the Gensim summarizer applied on a news article

3.3. Sumy LSA-Based Summarizer

Sumy is an implementation of various summarizers and the choice is given to choose from various summarization algorithms, the summarizer chosen for the comparison study in this thesis is latent semantic analyser (LSA)-based summarizer [70]. LSA is a fully automatic Mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in a passage of discourse [33]. LSA in the early works was a means to find corresponding documents in the search engines

from a given search word. This was a difficult task on its own given the complexity of the English language where an individual word depending on context may have different meanings, what LSA does in this case search for the concepts associated, doing this by mapping the words and the corresponding documents into a 'virtual' area and tries to map each word to a semantically accurate matching document. For this to work LSA algorithm is made in a way that simplifies the mapping process in the following way:

-Representing documents as a bag of words and no preference given to the order of appearance but only how often i.e. frequent a word appears.

-Representation of concepts by the pattern of frequently appearing together words for example the words "hostess", "plane", "ticket", "passenger" when they appear together indicate a concept and the document associated will be about aeroplanes/flights.

-for the above to work where words are mapped together to form a concept the number one assumption made will be that for any given word, that word only has one meaning and synonyms are assumed to be non-existent.

Summarization built on this knowledge where inputs were given as documents. The approach used in Sumy is through [33] which was inspired by latent semantic indexing to apply a singular value decomposition SVD to text summarization [33]. Stated that "SVD is capable of capturing and modelling interrelationships among terms and sentences." LSA is termed as an application of reduced-order SVD. The input is a document which is then represented as words (terms/features) vs documents or in cases sentences matrix. SVD then represents the matrix into singular vectors and their corresponding singular values allowing the documents or sentences to be summarized be represented in a virtual space which can be termed as 'latent semantic space'. When the documents or the sentences are placed in this space, they lead to creations of mappings which leads to placement of similar words and documents or sentences in the same area or together which happens even when they were originally not in the same area in the input document.

Singular Value Decomposition SVD:

This is an important concept in LSA based summarization because it forms the central base of the summarizer. SVD is a representation of input matrix into three individual vectors. It is said to “find a reduced dimensional representation of our input matrix that emphasises the strongest relationship and throws away the noise” [71] the best matrix is that which contains least possible information. SVD helps in this instance because by getting rid of the noise from the input document the stronger patterns get to be highlighted and hence the concepts are clearly aligned with their most relevant terminologies. From a mathematical point “SVD derives a mapping between the m-dimensional space specified by the weighted term-frequency vectors and the r-dimensional singular vector space, but from an NLP point of view this simply means that it derives “latent semantic structure of the document”. While normal algorithms capture word clusters SVD captures “semantic clusters” a magnitude of cluster is gotten based on how frequent it reoccurs and depending on this magnitude the most significant sentences can be scored based on how many strong concept it contains. The input text or document is transformed into matrix A which is later represented as three matrices U, S and V.

A: The Input Matrix:

The data is represented into number of documents versus the sentences in the M by N matrix where only the frequencies of the terms are taken into consideration.

U: Left Singular Matrix: An M by R matrix representing the documents or sentences and concepts into a space where they can correctly mapped into one another.

S: Singular Values

An R by R diagonal matrix which represents the score or the strength of each concept of the input text. R is the rank of the matrix A. this matrix has all the non-zero values represented in the diagonal in an order. The values present in the diagonal are termed as singular values they are arranged in decreasing order and determine the relative importance of the dimensions.

V: The Right Singular Matrix

A representation of the number of words which can be also the terms or features versus the concepts present in the input document.

The Matrices U and V are both a set of orthonormal vectors which means they are both orthogonal and normal. A set of vector is said to be orthogonal if any pair of vectors is orthogonal i.e. if and only if their product (equivalently cosine) is zero(0), while normality means that each singular vector is of length one (1).

From the above explanation of LSA and SVD we can briefly conclude that for all algorithms that use LSA for text summarization the steps taken are as shown in the steps below:

-Step 1: Representing the input documents in a set of sentences by term matrix. Each row representing a term where the term could mean a feature or a word, the columns represent the sentences. Values are given in terms of appearance of the term in a sentence where no appearance gives a score of zero and appearance gives a score of one for the times the term appears in the document. This is simply the frequency of the term in the corresponding sentence. In the example of the sentence “the food was as great as the service” the term the will have a score of two seeing that it appears twice in that sentence. If the word hate was in the rows it will get a score of zero in this particular sentences because it doesn’t appear at all in the sentence. This is the weight of the term in the sentence.

-Step 2: Application of SVD. The matrix created in step 1 I a sparse matrix containing a lot of zeros and occupying a lot of space, from the noise present in that matrix getting the correct mapping could be a hectic process, hence the representation of the input matrix into this smaller dimensional space model based on the frequencies given in step 1. The relations between the terms can therefore be better detected. Applied on text SVD is termed as LSA because it groups documents that are semantically related to each other even when they do not share common words [72].

From this point forward the different application of summarization algorithm pick the sentences with their particular strategies, in this thesis we discuss the one employed by Sumy which is an application from the original LSA for text summarization.

-Step 3: selection of the kth most right singular vector from matrix VT

-Step 4: select the sentences which has the largest index value with kth right singular vector and include it in summary.

-Step 5: the process is terminated if the limit given for the needed length is reached otherwise an increment by one is made and the process continue from three.

This is the method used where top sentences in top most concepts are picked by [33]. Another picked the same method but with slight improvement where they also considered the length of the sentence [73]. They considered sentences not only from the right singular value vector but also from the U matrix which had the ranks in [74]. However chose sentences based on three factors the right singular matrix VT, average value of each sentences and total length of each sentence in discovering concepts and sub concepts in [75].

The Sumy LSA summarizer was used in the thesis as part of the comparative study and the results are discussed in the result section. The figure shows the LSA-Based Summarizer.

Two powerful bombs exploded minutes apart outside the United States Embassies in Kenya and Tanzania this morning, killing at least 80 people, 8 of them Americans, in what officials said were coordinated terrorist attacks. In Nairobi, an enormous explosion ripped through downtown shortly after 10:30 A.M., turning the busy Haile Selassie Avenue into a scene of carnage and destruction that left more than 1,600 people wounded and dozens still missing long after night fell. …full article in [76]

Summary in 5 sentences.

At least eight Americans, one a child, and an unknown number of Kenyan employees of the embassy died in Nairobi in the blast which left the offices a honeycomb of burned-out rubble with bodies buried inside.

By nightfall, rescue workers and soldiers in Nairobi toiled under floodlights with backhoes to extricate dozens of bodies still buried in the rubble of the Ufundi House, which is situated behind the embassy.

American officials said that early circumstantial evidence was leading investigators to focus on a Saudi Arabian Islamic militant, Osama bin Laden.

Earlier this week the Islamic Holy War group, which is banned by Egypt, warned it would retaliate against Americans because of Washington's role in pressing for the extradition of three suspected terrorists from Albania to Cairo.

Out in front of the darkened embassy, American marines in full battle gear stood guard, wide-eyed and battle-bright, guns in combat position, as night fell.

Figure 3.4. Results of the Sumy LSA-Based summarizer applied on a news article

3.4. Sentiment Analysis Based Summarizer

For this summarizer and enhancement of the frequency based summarizer was implemented where sentiment analysis was first used to rule out the sentences that will not be included in the summary and then the frequency based summarizer was implemented in the remaining sentences. The argument for this summarizer was that in dealing with news articles most of the news items will either be termed as positive news which will have an inclusion of mostly positive sentences and in which case the negative sentences will not have an effect in the understanding the underlying context from summary hence the filtering out of the negative sentences in this scenario while the opposite applies to negative news article where we filter out positive sentences which will otherwise be just extra details with no real meaning to the news. The Steps for the algorithm applied is as below:

Step1: Get the text and cleaning of the text.

Step2: applying sentiment analysis to the text input and getting the positive and negative sentences respectively.

Step3: Depending on the sentences that are heavier (more in terms of the number of sentences) choose the relevant sentence to apply the summarizer on.

Step5: Apply the frequency based summarizer to get the summary required.

This summarizer was equally applied on online news article and given to users to evaluate the performance.

In order to understand the work done above there is need to understand what sentiment analysis is and how it was applied. The inspiration was taken from the different methods in which the fields in natural language processing interact, there is always knowledge borrowing among the different fields where successful algorithms in a field is borrowed and applied to another field sometimes with a variation but sometimes with the same steps and getting great results. Sentiment analysis has been applied together with text summarization before where integrated the sentiment analysis with text summaries to get the sentiment of the particular summary [60]. While use of sentiment analysis where the sentences were ranked according to importance in [27]. Sentiment analysis as a field on its own is also referred to us opinion mining, it is a study that mainly focuses on getting and analysing opinion sentiments evaluating appraisals attitudes and emotions [77]. We have three levels of analysis when it comes to sentiment analysis: at the document level i.e. classifying whether a whole document is negative or positive, the sentence level checks for particular sentence and whether it is negative, positive or neutral while the entity and aspect level returns finer-grained results which is not possible with the other two, it gets exactly what people like or don't like. To get sentiment analysis mostly classifiers are trained to identify which sentences are positive and which ones are not, some of the most popular algorithms in sentiment analysis are briefly shown below:

Logistic Regression:

It is a technique borrowed by machine learning from the field of statistics. It is named for the function used at the core of the method, the logistic function. It involves a linear

discriminant giving a probability that given input point belongs to a certain class. This classifier deals with linear separable data i.e. categories.

Naïve Bayes:

This is one of the most popular algorithm used in text classification, it's relatively faster compared to other algorithms and can be used as a quick fix to classification. It works on Bayes theorem of probability to predict the class of unknown data set based on its features, Naïve Bayes is easy to train, understand and is not sensitive to irrelevant features, its main disadvantage is that it assumes every feature is independent which normally isn't the case. Naïve Bayes creates a likelihood of whether a text is positive or negative based on the features of the given text.

Multinomial Naive Bayes:

Implements the naive Bayes algorithm for multinomially distributed data.it is used for discrete counts, it gives the classification as per "number of times outcome number x_i is observed over the n trials".

Bernoulli Naïve Bayes: This classifier is based on data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, Boolean) variable hence it requires features to be binary valued. This classifier is different from multinomial Naïve Bayes in that it penalizes non-occurrence of a feature while multinomial simply ignore a non-occurring feature, it works well with short texts.

Linear Support Vector Classification: Fits the data provided i.e. training data returning a "best" hyper plane that divides or categorizes data, then according to the test data gives the predicted category.

In the implementation of the sentiment analysis section of this summarizer the classifier used in training of identification of the positive from negative sentences was the Logistic Regression Classifier. The training was done on a dataset comprising of 1000 positive and 1000 negative movie reviews. This dataset was retrieved from [78]. Which was created for the purpose of natural language processing by [79]. After the cleaning of the dataset and splitting to training and test set the classifier was applied

on the data and later saved as a module together with the model that tfidf model that transforms the data into the same format as that which was officially trained. This classifier was called as a module and the applied on news article which gave a negative or positive classification of the input text sentences aiding in the division of the text to give a reduced sentences to start with in the summary giving room for better efficiency. The summarizer was also tested on the users after its application on a news article and the result section shows the outcome.

Read full coverage of the royal wedding, continually updated by CNN reporters. London (CNN) -- Prince William of Wales slipped a gold ring onto the finger of Catherine Middleton Friday, and the couple vowed to love, comfort, honor and to keep each other in London's biggest royal wedding in three decades. Bells pealed over central London and flag-waving crowds roared in excitement Friday as Middleton arrived at Westminster Abbey to marry William, the second in line to the British throne. Middleton wore an ivory and white satin dress with lace sleeves and shoulders, designed by Sarah Burton of the Alexander McQueen fashion house. Royal wedding: The big day "You look beautiful," the prince told her as she arrived at the altar on the arm of her father Michael. William wore the uniform of a colonel of the Irish Guards, a scarlet jacket and blue sash, as his brother, Prince Harry, accompanied him into the abbey. Crowds cheered as his car drove the short distance from Clarence House to the abbey before the wedding, and they roared and waved as the newlyweds rode in an open carriage from the abbey to Buckingham Palace after the ceremony…full [80]

Five sentences summary gives:

Former Prime Ministers Tony Blair and Gordon Brown were not invited, leading to accusations that the royal family favours the Conservative party over Labour.

They met as college students at the University of St. Andrews in Scotland, sharing an apartment with a circle of friends before they began dating.

The royal family explained that as a matter of protocol, presidents were not invited.

Read full coverage of the royal wedding, continually updated by CNN reporters.

British Culture Secretary Jeremy Hunt predicted the ceremony would be seen by an estimated 2 billion people worldwide.

Figure 3.5.Results of the Sentiment Analyser-Based summarizer on a news article

4. EVALUATION MEASURES AND RESULT ACHIEVED

This section covers the evaluation of the summarization and the results achieved from the comparative study. The evaluation measure for text summarization are discussed exploring the different methods employed in evaluating the summarizers. Then the evaluation method used in the thesis study is discussed thereafter giving the results of the study. This section is completed by a comparison between Rouge Evaluation and the proposed evaluation method termed as usability study.

4.1. Evaluation Measures for Text Summarization

With all the efforts made in automating text summarization it is a positive change to work towards automation of evaluation of the work done in text summarization, however since as it is seen the whole effort is done for human use the argument here is that it is only right if we completely do not overlook the input of humans in the evaluation section of summaries and summarizers. Currently there are different evaluation techniques in existence, this depends on whether it's being compared to the source text, another automatically generated summary or to a human generated summary, and in its basic form evaluation techniques are divide into two types and then further divided in subsequent categories [81].

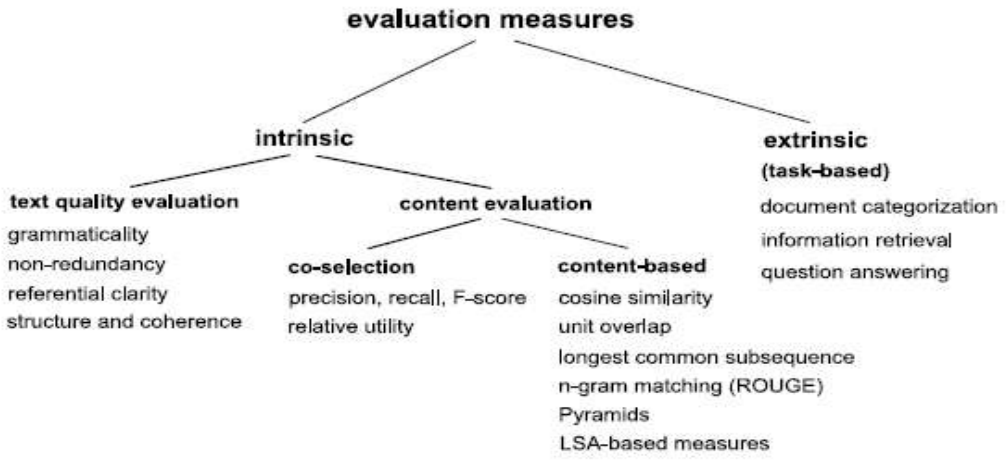


Figure 4.1. Existing evaluation measure [81]

Extrinsic: Checks on how helpful a given summary is for a particular given text

Intrinsic: Directly based on analysis on analysis of the summary done by either checking in comparison to source text or the abstract by human, this method can be categorized into content and text quality, all the methods are explained below in detail:

4.1.1. Text quality measures based methods

This measure are a face-value evaluation of the automatic summary produced and its one of the earliest evaluation technique applied not automatically but by human judges who assign marks ranging from A- very good to E – very poor to a given summary according to how well it comes out expressing the text quality and information given this methods are as discussed below:

Grammar: The given summary is checked for whether it contains any grammatical error in Basic English or the language used, this errors include, punctuation, incorrect word inclusion, non-textual terms etc., the evaluated automatic summary is categorized as either a decent or mediocre summary based on the point given grammatically.

Non-Redundancy: the main point in getting a summary is to reduce the information to a readable informative size hence the red judgement in inclusion of redundant word or sentences, the less repetition of the text in the summary the better the summary evaluation summary.

Reference Clarity: A summary that has a proper referential content in terms of then features like pronouns and nouns etc. give information that makes more sense to the reader and hence evaluated better than a summary that mixed up all the information losing the point or the reference person or thing in the process.

Coherent and Structure: If the text produced as summary makes sense to the user in terms of coherence and the sentence structure its highly marked as a good summary.

Content evaluation measures based methods

This methods are an improvement from the previous evaluation methods because they give a level of automation which eliminates the use emotionally- biased human

evaluators, this methods are categorically divided into two: Co-selective- and content-Based measures which are described in brief detail below:

Co-selective Measures:

This measures check for the effectiveness of a summary as a comparison to the system summary i.e. the automatic summary and ideal summary and there are three ways to do the co-selection analysis:

a) Precision , Recall and F-Score:

Precision Measure: a measure that combines the number of sentences in summaries from both the automatic system and ideal summary and then divides it by the number of sentence in the automatic system summary.

Recall Measure: This measure is similar to the precision measure in the fact that it combines the sentence from both the automatic and ideal summary sentences but unlike the former this one divides by the number of sentence in the ideal summary.

F-Score: This measure is a composite measure combining the two above mentioned measure, the Recall and Precision measure to get the effectiveness of a summary.

b) Relative Utility:

In the previously discussed method of evaluation, the precision and recall measures may be too harsh or biased in terms of judging summaries as the ideal summary is constructed by human judges and hence may at times judge two similar summaries in a different score measure, Relative utility was as solution to the problem in that a utility score is assigned to each sentence in a document based on the degree of relevance of that particular sentence I.e. how informative a particular sentence is, this utility scores are given by human judges , based on this values the effectiveness of a summary is calculated by checking the value of the sentences it included in its summary.

Content Based Measures:

The co-selective measures are also superficial and don’t delve deep into meaning to check the effectiveness of a summary but only based on the sentence match, content

based measures on the other hand focus on the content of the summary and can get the similarity of two separate sentences based on their informativeness. This measures are:

Cosine Similarity:

Most basic form of content similarity measure and uses vector space model to gauge the similarity between a summary and its reference document

Unit Overlap:

Gets the measure by checking the overlap between the summary and the reference document by use of sets of words or lemma.

Longest common subsequence:

This measure judges a summary in terms of the length of the longest common subsequence based on words or lemmas between the summary and the reference document.

N-gram Co-occurrence Statistics –ROUGE:

In the quest to get a fully automated evaluation measure ROUGE (Recall-Oriented Understudy for Gisting Evaluation) was introduced in the Document Understanding Conference 2003, this evaluation measure is based on similarities of n-grams between the auto created summary and a reference summary. Based on the granulite evaluation needed there are different ROUGE scores , the reference ROUGE measure is the ROUGE-N measure gets the n-gram measure, other scores are ROUGE-L i.e. Longest subsequent ROUGE measure and ROUGE-SU4 a bigram measure that allows for as high as four unigrams. N-gram simply means a set of a co-occurring words in a predefined length of N for example given the sentence “the girl was over the moon with happiness” if n=2 then we have “the girl” “girl was” “was over” “over the” “the moon” “moon with” “with happiness” giving as a set of 7 n-grams and so on[82].

Pyramid Measure:

This is a semi-automatic evaluation measure that is based on identifying summarization content unit (SCU’s) and arranging them in form of a pyramid, the

SCU's with the highest value are placed on top of the pyramid, the pyramid from extracted summary is pitted against the pyramid for the manual summary to gauge similarity and strength of the automatic summary.

LSA Based Measure:

This method contributed evaluated the quality of summary by checking for the content similarity of the summary and the source document by [73]. This was done by use of the singular value decomposition (SVD) matrix which was discussed in the LSA based summarizer section. By deriving a U matrix of both the source and summary document this method compared the similarity between the two to get the summary quality.

4.1.2. Extrinsic (task based measures)

This measures unlike the previous intrinsic measures that made comparisons between an ideal summary and the candidate summary check for how effective a summary is based on how it can deal with a real world task at hand instead. This measures take the summary produced and check how effectively they solve the user's need in a particular task, the most used or common task based measures are discussed in the following region:

Document Categorization:

Evaluates a summary by how based solely on the summary a document can be classified according to the category it belongs to given a choice of particular topic categories, this could be done either manually or by use of machine but using a machine could give room to inclusion of an automatic machine classification error, therefore the best measure is the human evaluation, and by use of precision, recall and f-score the score of the summary in efficiently categorizing a document classifies it as either a good or a bad summary.

Information Retrieval:

As a substitute for the source document, can the summary capture the main subject and points in the reference document and be used purpose of information retrieval? If the summary is itself sufficient for such a task then it is evaluated as a good summary.

Question Answering:

The extracted summary has to be sufficient in itself to answer subsequent answers taken from the reference document, test subjects are tested on how well they can answer questions from a particular subject given only the generated summary and the summaries are evaluated accordingly.

In conclusion, the method of summarization evaluation depends on the level of technology, purpose (goal) and time required for evaluation. The favoured evaluation is fully automated evaluation which would eliminate any outside bias whether from machines or from human evaluators.

All the explained methods above are already existing both human and automatic evaluation methods, in the thesis study however we borrow a method from Human computer interaction and introduce in text summarization a method that has been popular in the field of HCI. This method is termed as a usability test and it is discussed below:

4.1.3. Proposed evaluation metrics-usability study evaluation

For the evaluation of the systems to get a comparative study of the different summarizers, a popular evaluation system was borrowed from the field of HCI, placed under the task based evaluation measures, the proposed metrics chose 10 participants to take a usability test of the summarizers. The participant were charged with going through a list of given tasks, taking a note of whether they managed to do the tasks and after that were requested to fill in a system usability scale(SUS) hereby referred to as summarizer usability scale. The SUS system is an adaptation of the original scale introduce by John Brooks in 1986, in that the questions are sequenced just like the original scale with a "quick and dirty" evaluation based on ten questions, although similar in pattern the question were edited for the purpose of the study hence the term 'adoption'[83].

This scale works by first asking the users to use the system for given tasks and based on their interactions with the system they give the score. The tasks (attached in the appendix) took a record of the difference in reading timing of the summary and article

respectively, users were asked question in regards to the summary. The users were then tasked with assigning scores of the system based on 10 heuristics of system usability, the scale has ten question which includes interchanging negative and positive aspects of the summary and for each summary based on the tasks undertaken they were asked to give a score of 1 to 5 going from highly disagree to highly agree respectively. In addition to English comprehension it was decided to include both fast and slow readers in order to really get the impact of time difference in article and summary reading time. The users were reminded that the task was volunteer based and although completion of the tasks was a requirement it was not a must and could leave the task incomplete if necessary as is needed to be done in all user based studies, in addition users were encouraged to think aloud and ask questions where not sure what is needed. The table below shows the details of the demographics of the participants used:

Table 4.1. Demographics of the users in usability study

Users	Level of Study	Type of Reader
User 1	PHD	Fast
User 2	Undergraduate	Fast
User 3	Masters	Medium
User 4	PHD	Slow
User 5	Masters	Slow
User 6	Masters	Medium
User 7	Undergraduate	Medium
User 8	Undergraduate	Slow
User 9	Masters	Medium
User 10	Masters	Fast

"An investigation into the cause, manner and circumstances of the death is currently underway by the Investigations and Coroner Divisions of the Sheriff's Office," the sheriff's statement said.

"Word of Williams' death stunned the entertainment community and beyond Monday. President Barack Obama's statement sent from the White House summed it up: "Robin Williams was an airman, a doctor, a genie, a nanny, a president, a professor, a bangarang Peter Pan, and everything in between.

"Coroner investigators suspect "the death to be a suicide due to asphyxia," according to a statement from the Marin County, California, Sheriff's Office. Williams married graphic designer Susan Schneider in Napa Valley, California, ceremony in October 2011.

(CNN) -- Robin Williams -- who first made America laugh and eventually touched "every element of the human spirit" in a remarkable range of performances -- died at his Northern California home Monday. Williams apparently took his own life, law enforcement officials said.

""Happy Days" star Henry Winkler said it was "unimaginable that this is the reality today, that this incredible human being, incredible, delicate, funny, dramatic human being is gone.

Figure 4.2. Summary provided for Simple Summarizer Usability Study

After the users signed a consent form, they were instructed to first read the respective summaries generated by the different algorithms of an online article. Different articles were used for the different summarizers a total of four articles across four summarizers and ten user.

But the opposition leader, Raila Odinga, challenged the results, and the Supreme Court nullified the election in September, citing irregularities. Backers of Mr. Kenyatta interpreted both of his wins as broad national support for the president, but opposition supporters said they had twice been disenfranchised by a process that lacked credibility. Mr. Odinga withdrew from the second election two weeks before the vote, arguing that the electoral commission could not oversee a free and fair process, and he called on his supporters to boycott.

The top elections official, Wafula Chebukati, warned a week before the polls opened that political interference in the commission’s work was likely to undermine the credibility and neutrality of the vote. Mr. Chebukati backtracked on that criticism while announcing the results on Monday, declaring the process “free and fair.” At least 14 people have been killed in election-related violence since the Oct. 26 vote, according to international officials, and more have been injured.

“The killing was well planned and executed,” he said, and was intended as “a direct warning to others.” Mr. Odinga condemned the violence, which he described as perpetrated by Kenyatta supporters, but stopped short of urging his own supporters not to react with violence. Kenyans, foreign diplomats and even some of Mr. Kenyatta’s own supporters have called on the president to engage with opposition leaders after his win, but in his victory speech the president distanced himself from dialogue. “Those who want to ask me, ‘Are you going to engage in dialogue with so and so and so and so?’ let them first and foremost exhaust the constitutionally laid out processes,” Mr. Kenyatta said.

Figure 4.3. Summary provided for the Gensim Summarizer Usability Study

At least eight Americans, one a child, and an unknown number of Kenyan employees of the embassy died in Nairobi in the blast which left the offices a honeycomb of burned-out rubble with bodies buried inside.

By nightfall, rescue workers and soldiers in Nairobi toiled under floodlights with backhoes to extricate dozens of bodies still buried in the rubble of the Ufundi House, which is situated behind the embassy.

American officials said that early circumstantial evidence was leading investigators to focus on a Saudi Arabian Islamic militant, Osama bin Laden.

Earlier this week the Islamic Holy War group, which is banned by Egypt, warned it would retaliate against Americans because of Washington's role in pressing for the extradition of three suspected terrorists from Albania to Cairo.

Out in front of the darkened embassy, American marines in full battle gear stood guard, wide-eyed and battle-bright, guns in combat position, as night fell.

Figure 4.4. Summary provided for LSA summarizer usability study

Former Prime Ministers Tony Blair and Gordon Brown were not invited, leading to accusations that the royal family favours the Conservative party over Labour.

They met as college students at the University of St. Andrews in Scotland, sharing an apartment with a circle of friends before they began dating.

The royal family explained that as a matter of protocol, presidents were not invited.

Read full coverage of the royal wedding, continually updated by CNN reporters.

British Culture Secretary Jeremy Hunt predicted the ceremony would be seen by an estimated 2 billion people worldwide.

Figure 4.5. Summary provided for Sentiment analysis usability study

After reading the summary, for every given summarizers users were asked to answer the questions below:

How long did it take you to read the summary?

What is the article about based on just the summary?

What is the event category based on the summary (events are happening i.e. Death, Terror etc.)

Does the summary give you enough idea on whether you want to read the full article for details?

Does the summary affect your attitude towards the whole article?

After completing this the users were asked to read the source article and answer the questions that followed:

How long did it take you to read the source article?

Did the article contain extra information to improve on your understanding of the context of the article?

Did reading the article after the summary, change your opinion on the story?

After the completion of the questions which were termed as the task the users were requested to complete a system scaling form that contained ten question in regard to how they ranked the summarizers. This questions were adopted from the format of the original system usability scale used in HCI. They were tailored to suit the evaluation of summarizers based on the grammar, length, time saved, the comprehension capability of summaries and the flow of the summary. Users then rated from strongly disagree to strongly agree a scale of 1 to 5. The questions are shown in the form:

- I think that I would like to use this summarizer frequently.
- I found the length of the summary not satisfactory
- I could comprehend the story just from summary.
- I thought I needed full article to understand the story.
- I found the sentences coherent and with good flow.
- I thought there was discontinuation of the information given.
- I would imagine that most people would understand the story quickly.
- I found the summary sentences very grammatically inconsistent.
- I felt I saved time using the summary.
- I found no change between time taken to read summary and the actual article

This questions are termed as ‘quick and dirty’ method of getting the quality of a system. And the summarizer being a system that is intended for the use of the public use was fit to undergo the test. The results showed that with the correct number of users the evaluation method introduced could be used as a metrics for summary evaluation. The results of the usability study is discussed in the results section and also there is a comparison between the proposed evaluation metric and the Rouge evaluator.

4.1.4. Rouge evaluation metrics

The metrics already discussed in the section above, usually recognized as the standard evaluation metric was used to get a comparative evaluation of the summaries to check against the usability study used in the evaluation of the summarizers. The Rouge calculation used was taken from an implementation of Java termed as Rouge 2.0 that is used for the same as the original Rouge implemented in Perl Language. The main component of a Rouge evaluation is getting the system and reference summary. A system summary is the summary generated by the algorithms i.e. the automatically produced summary, while the reference summary refers to a gold standard summary which could be one or more summary usually generated by human judges. To generate the system summary for the evaluation of the four summarizers, five random online articles were used, the same were applied for all the summarizers to have uniform results. The articles were run through the summarizers and the respective summaries produced were named according to the format that is specified to work with Rouge2.0. The summaries were named as newsX_syssum1, syssum2, syssum3, syssym4 for the Simple, Gensim, LSA-based and Sentiment Analysis summarizers respectively, X standing for the specific news article numbered for ease of identification i.e. first news article is named news1_syssum1 for simple summarizer. For the articles human produced gold standard or reference summaries, human judges were asked to pick the sentences they termed as the most important for the story to make sense i.e. the most influential sentences. Since the summarizers implemented and extractive based method the same was maintained for the reference summaries to avoid inconsistencies. The sentences picked by at least two human judges were chosen and taken as they were from the source article and stored as dictated by Rouge 2.0 documentation as newsX_reference1 maintaining the same naming procedure as of the system produced summaries.

The summaries generated by the system was compared to the golden summary to get effective Rouge-1 and Rouge-2 scores. A sample reference and system generated summaries are shown in the following figures.

An investigation into the cause, manner and circumstances of the death is currently underway by the Investigations and Coroner Divisions of the Sheriff's Office, the sheriff's statement said.

Word of Williams' death stunned the entertainment community and beyond Monday. President Barack Obama's statement sent from the White House summed it up: Robin Williams was an airman, a doctor, a genie, a nanny, a president, a professor, a bangarang Peter Pan, and everything in between.

Coroner investigators suspect the death to be a suicide due to asphyxia, according to a statement from the Marin County, California, Sheriff's Office. Williams married graphic designer Susan Schneider in Napa Valley, California, ceremony in October 2011.

(CNN) -- Robin Williams -- who first made America laugh and eventually touched every element of the human spirit in a remarkable range of performances -- died at his Northern California home Monday. Williams apparently took his own life, law enforcement officials said.

Coroner Division suspects the death to be a suicide due to asphyxia, but a comprehensive investigation must be completed before a final determination is made.

Williams, born in Chicago on July 21, 1951, studied theatre at Juilliard School before taking his stand-up act to nightclubs.

Figure 4.6. Reference summary used for Rouge Evaluation

An investigation into the cause, manner and circumstances of the death is currently underway by the Investigations and Coroner Divisions of the Sheriff's Office, the sheriff's statement said.

Word of Williams' death stunned the entertainment community and beyond Monday. President Barack Obama's statement sent from the White House summed it up: Robin Williams was an airman, a doctor, a genie, a nanny, a president, a professor, a bangarang Peter Pan, and everything in between.

Coroner investigators suspect the death to be a suicide due to asphyxia, according to a statement from the Marin County, California, Sheriff's Office. Williams married graphic designer Susan Schneider in Napa Valley, California, ceremony in October 2011.

(CNN) -- Robin Williams -- who first made America laugh and eventually touched every element of the human spirit in a remarkable range of performances -- died at his Northern California home Monday. Williams apparently took his own life, law enforcement officials said.

Happy Days star Henry Winkler said it was unimaginable that this is the reality today, that this incredible human being, incredible, delicate, funny, dramatic human being is gone.

Figure 4.7. Simple summarizer system summary (news1_syssum1)

(CNN) -- Robin Williams -- who first made America laugh and eventually touched every element of the human spirit in a remarkable range of performances -- died at his Northern California home Monday.

Williams apparently took his own life, law enforcement officials said.

The family respectfully asks for their privacy as they grieve during this very difficult time.Coroner investigators suspect the death to be a suicide due to asphyxia, according to a statement from the Marin County, California, Sheriff's Office.

Williams married graphic designer Susan Schneider in Napa Valley, California, ceremony in October 2011.This morning, I lost my husband and my best friend, while the world lost one of its most beloved artists and beautiful human beings.

He gave his immeasurable talent freely and generously to those who needed it most -- from our troops stationed abroad to the marginalized on our own streets.Comedian Steve Martin tweeted, I could not be more stunned by the loss of Robin Williams, mensch, great talent, acting partner, genuine soul.Former CNN host Larry King said he would remember Williams as a genuine caring guy.

Figure 4.8. Gensim System Summary (news1_syssum2)

(CNN) -- Robin Williams -- who first made America laugh and eventually touched every element of the human spirit in a remarkable range of performances -- died at his Northern California home Monday.Williams apparently took his own life, law enforcement officials said.

Comedian Steve Martin tweeted, I could not be more stunned by the loss of Robin Williams, mensch, great talent, acting partner, genuine soul.

An autopsy is scheduled for Tuesday, the sheriff said.Williams made at least two trips to rehab for drug treatment, including a visit this summer, and he underwent heart surgery in 2009.Williams, born in Chicago on July 21, 1951, studied theater at Juilliard School before taking his stand up act to nightclubs.

The role led to the spin-off show Mork & Mindy, which showcased Williams' usual comic improvisation talents.He proved his dramatic acting skills in Good Will Hunting, a 1997 film that earned him a best supporting actor Oscar.His memorable movies over the past three decades includes Good Morning, Vietnam, Dead Poets Society, Mrs. Doubtfire and The Birdcage.

Williams' fans can look forward to four more movie appearances coming to theaters, including another installment in the Night at the Museum franchise.The film, set for a December release, has Williams reprising the Teddy Roosevelt role he delivered in the first two comedies.Share your memories of Robin WilliamsSee more comedy content at CNN Comedy.CNN's Travis Sattiewhite, Rachel Wells and Carolyn Sung contributed to this report.

Figure 4.9. LSA System Summary (news1_syssum3)

This morning, I lost my husband and my best friend, while the world lost one of its most beloved artists and beautiful human beings.

The list is much longer. Williams credited the influence of Jonathan Winters' comic irreverence and quirky characters as a great influence on his comedy.

He arrived in our lives as an alien -- but he ended up touching every element of the human spirit.

Happy Days star Henry Winkler said it was unimaginable that this is the reality today, that this incredible human being, incredible, delicate, funny, dramatic human being is gone.

When Winters died in 2013, Williams said he was my idol, then he was my mentor and amazing friend.

Figure 4.10. Sentiment-Based system Summary (news1_syssum4)

For all the system generated summaries corresponding Rouge scores was calculated. The summary used as the reference was only one summary. The result of the Rouge score is discussed in the result section and thereafter comparison between the two evaluation metrics is done.

4.2. Findings

In this section we discuss the findings based on the evaluation of the four systems. The result section is divided into four sections; the first section discusses the finding from the usability study task evaluation where users were asked to do certain task and answer corresponding questions, and the second section will be a discussion from the findings of the summary evaluation scale where the adaptation of the system usability scale for the summarizer ranking is discussed. The third section will discuss the Rouge findings from the Rouge score done on the four comparative summarizers and the final section will discuss the comparison between Rouge and Usability Study evaluation metrics.

4.2.1. Usability study task evaluation

As discussed in the evaluation measures, this proposed evaluation metrics from the field of HCI entailed giving the participants a list of tasks to do. The task included reading the summary and then answering questions based on the understanding of the storyline and after that reading the source article and answering a few more question on whether the understanding of the storyline had changed, the timings of reading the summary and source articled were also recorded. This was repeated for all the summarizers and the results depending on the factors tested are as follows:

Time Factor: For all the systems the times were recorded for how long it took to read an article and the summary and this time factor was depended upon when deciding how beneficial a summarizer is, for this metrics the deciding factors were whether indeed without reading the article the user could get the story and hence won't need to read the full article hence saving more than triple the time and the other factor was whether the summary was enough to make a decision on whether the article is important enough to continue to the full story hence saving time by the user just reading articles that are of interest to them or that have raised enough curiosity to be a deciding factor.

The time taken to read the article was as expected more than the summary with at times, depending on the summarizer, the time taken to read the article taking more than five times the time to read its respective summary. This measure is useful in our argument that indeed the user saves time in reading content without wasting as much time as it would have otherwise, however this measurement alone cannot measure usefulness of a summary because a summary may take less time to read, as expected of summaries but how much of the information given was useful and cancelled the need for reading the whole document? This is where the differences in the summarizers were first noted. For the simple summarizer, the users pointed out that indeed they saved time because not only did they get the concept but they also did not find the reading of the article any more informative i.e. they could stay satisfied with just the output of the summary and this in fact answers one of thesis argument that indeed for the average user the time taken to read an article is massively reduced by reading an article summary by means of elimination of reading the article or in deciding which article is worth the time. For the Gensim system the result of saving time was less strongly portrayed because most users felt the details of the article was necessary for the understanding of the storyline but at the same time the users when asked whether the summary could affect whether they want to read the article or not they agreed that from the summary they could choose whether the article is worth reading, the same was for LSA summarizer but for the sentiment analyser three participants felt they didn't save anytime in terms of the summary not being beneficial i.e. just the summary only was not enough to help in decision making, although the majority agreed that their curiosity to read the article was affected by the summary.

Storyline and Event: the users were given the summary to read and from that were asked whether they could tell the event in the story and the storyline without getting the source story. It was assumed here that all the users had not read the story before. For the simple summarizer all the participants could tell the storyline and the event corresponding to the story, the same for the LSA summarizer.

In Gensim summarizer however, one participant could not pinpoint the event in question even though they got the storyline stating that they could not pick just one category, for the sentiment summarizer the seven out of the ten participants could give the story line the rest were not sure what the story was about and eight of the users could tell what events in the story, one participant could say the event but did not know what the storyline was.

Curiosity Measure: This is the measure of how much a user wanted to proceed to reading the article after reading the summary for all the summarizers more than half the participants claimed the summary affected their choice on whether or not to read the article but for the LSA summarizer the participants were split in half with one half saying the summary was enough to measure their curiosity meter while the other half said the summary did not give them enough to go with in order to decide.

Attitude Measure: the participants were asked whether the summaries affected their attitudes towards the whole article and whether they were able to form and opinion on the matter in discussion, for the simple summarizer it was a split decision the users said they could get the story and whether to read or not but their opinion could not be fully affected unless they read the article for comparison. The Gensim summarizer got a more yes answers with the eight users stating that they had their opinions from summary. The other two summarizers had equal answers with six users stating that they had formed an attitude towards the article from the summary.

After the tasks above the users were requested to read the full article and their responses to the remaining questions were recorded as below:

Improved Understanding: the users were asked whether the information contained in the full article improved their understanding which will otherwise not have made sense. For the simple summarizer, the majority of the participants (six) stated the fact

that the information contained was extra details that was not essential to the understanding of the general story. For the Gensim summarizer the ration was split in half with some user stating the article contained information that was essential in the understanding of the storyline while the other half maintained that indeed the article had extra information though one could still get the main idea of the story without it.

For the other two summarizers; LSA and Sentiment, majority of the participants (8 and six respectively) claimed the extra information contained in the article was a major detail toward the understanding of the full story.

Opinion change: The users were questioned on whether their opinions on the story changed after reading the full article only three users felt that their opinion changed after reading the summary from the simple summarizer and the same was shared by the Gensim summarizer. Though the LSA analyser was split in half the majority of sentiment analyser claim (six) was that the opinion changed on the perspective of the given story.

4.2.2. Summarizer usability scale

After the completion of the tasks, the participants were asked to give an equivalent system usability scale only in this thesis the questions were customized to fit the testing of a summarizer, the questions were centred on the length, grammar, comprehension and time saved. For the grammar analysis the question on whether the sentences were correct grammatically was met with positive affirmation, this was expected given that the summaries were based on extractive methods which just took the sentences as they were in the source article with no paraphrasing. The other measures are as straight forward as the questions given in the questionnaire provided as an attachment with the exception on the time saved element which was explained to the user that it did not mean time taken to read article versus the summary but rather time saved in terms of decision making factor on whether to read full article or not and whether the summary was enough to understand the story. The results of the SUS scale is given below, but before that an explanation of how SUS works is give (84):

A SUS score of 80.3 and above indicates an excellent score with users having a higher chance of using the summarizer and recommending to friends. A score in the ranges

of 68 is an average score with indication of an okay summarizer but with a need for improvement. A score with less than 51 indicated a poor system with need for immediate fixing. Getting the findings is a bit hectic with lots of mathematic involved however an existing excel sheet calculation method explained in detail could get the results in an easier way, the excel sheet is comprised of four parts each for the individual summarizers, it is attached in the file together with the project by [84].

The results are given in the table below showing the highest score given for any particular summarizer in bold also the highest average highest score is given in bold, after that the results are discussed in details according to each summarizer :

Table 4.2. Individual and average scores given for the summarizers

Users	Simple Summarizer	Gensim Summarizer	LSA summarizer	Sentiment Summarizer
user 1	77,5	**70**	**77,5**	37,5
user 2	72,5	32,5	50	32,5
user 3	72,5	50	57,5	60
user 4	80	52,5	**77,5**	25
user 5	80	**70**	70	65
user 6	**95**	25	**77,5**	**77,5**
user 7	80	**70**	67,5	75
user 8	85	65	55	75
user 9	92,5	**70**	75	75
user 10	77,5	62,5	52,5	70
Average Score	**81,25**	56,75	66	59,25

Simple Summarizer:

The summarizer was given above excellent by three users and five users giving a score of close to excellent while the rest two gave a slightly above average score. In total the

average score given for this summarizer was 81,25 which is close to an excellent score and in fact be categorized as excellent.

This results reflects the opinion of the users as they went about doing the tasks given, most users had been seen as earlier discussed to get the idea of the story and pointed to the fact that the summary was indeed enough hence no need for the summary.

Gensim Summarizer:

Four people gave a score of above average with one person giving close to average another slightly above poor and the rest four giving below poor, the system generally got an average of 56,75 which translates to slightly above poor and needs to be improved. In comparison to the other summaries this summariser came in as the last choice among the summarizers, the users cited a lack of proper comprehension of the story line.

LSA Summarizer:

Four users gave above average score while the rest of the users were split between close to average ranges, above poor and below poor respectively, with an average for the system of 66 giving a close to average system with need for improvement. This summarizers was an average summarizers, the users opinionated that the summary was okay to get a general idea of the story and to make a decision on whether or not they want to read the full article but it lacked short in some aspect like leaving out some of the most important information, without which the story would not have been completely understood.

Sentiment Summarizer:

Only 1 person gave a score close to excellent while four found it average, 2 between poor and average score and 3 very much under poor performance. The average for this summarizer was 59,25 slightly above poor but less than average which indicates a need for improvement. The summarizer did slightly better than the Gensim summarizer, the users cited lack of detailed information as the reason for such small score but other than that as a guide on whether or not the story was worth the time it could be relied on.

From the summary of the results above and user feedback the best summarizer was found to be the simple summarizer with the LSA based following with less point, then the sentiment summarizer followed closely by Gensim summarizer. The results indicated that the summarizers had all been chosen by different users according to the summarizer they would like to continue using and all of them had at least one user giving an above average score meaning the summarizer could be the summarizer of choice. The users had also pointed out that all the summaries could be used as a quick guide on whether or not to proceed with reading the full article, but the difference came in whether the summarizer could be used as a standalone in case the user need a quick gist of the story.

4.2.3. Rouge results

In calculating the Rouge score, we used Java programming language implementation of the original Perl Programming Language version of Rouge that has gained popularity in usage. There are three measures that are usually calculated to get the effectiveness of a summarizer. The computed measures are Precision (P-Score), Recall(R-Score) and F measure. Recall score calculates how much of the reference summary is captured by the generated system summary, this simply translates to a division of the total number of overlapping words between the two summaries by the total words in the reference summary. However there is room for bias where the created summary might be a result of a long system generated summary that has all the words in the reference summary but is in fact a weak summary, this is where precision comes in, a calculation of how much of the overlapping are efficient by making division of the overlapping words by the total words in the summary. This may also penalise the system. F-score is the harmonising factor of the two P and R scores.

Table 4.3. Sample Rouge-1 Scores

ROUGE-Type	Task Name	System Name	Avg_Recall	Avg_Precision	Avg_F-Score
ROUGE-1	NEWS1	SYSSUM1.TXT	**0,78818**	**0,85561**	**0,82051**
ROUGE-1	NEWS1	SYSSUM2.TXT	0,51232	0,56216	0,53608
ROUGE-1	NEWS1	SYSSUM3.TXT	0,50739	0,4187	0,4588
ROUGE-1	NEWS1	SYSSUM4.TXT	0,17241	0,30435	0,22013

Table 4.4. Sample Rouge -2 Scores

ROUGE-Type	Task Name	System Name	Avg_Recall	Avg_Precision	Avg_F-Score
ROUGE-2	NEWS1	SYSSUM1.TXT	**0,78173**	**0,85083**	**0,81481**
ROUGE-2	NEWS1	SYSSUM2.TXT	0,40102	0,43889	0,4191
ROUGE-2	NEWS1	SYSSUM3.TXT	0,30457	0,24896	0,27397
ROUGE-2	NEWS1	SYSSUM4.TXT	0,02538	0,04545	0,03257

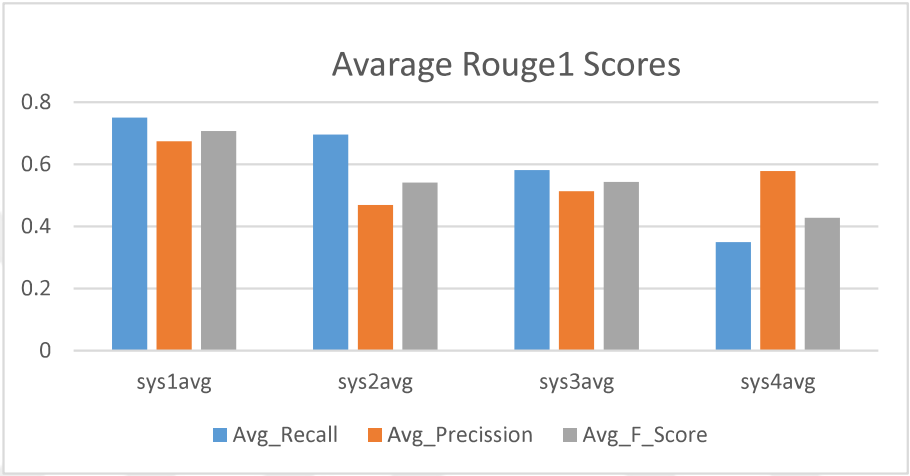


Figure 4.11. Chart comparison of the average Rouge-1 Scores

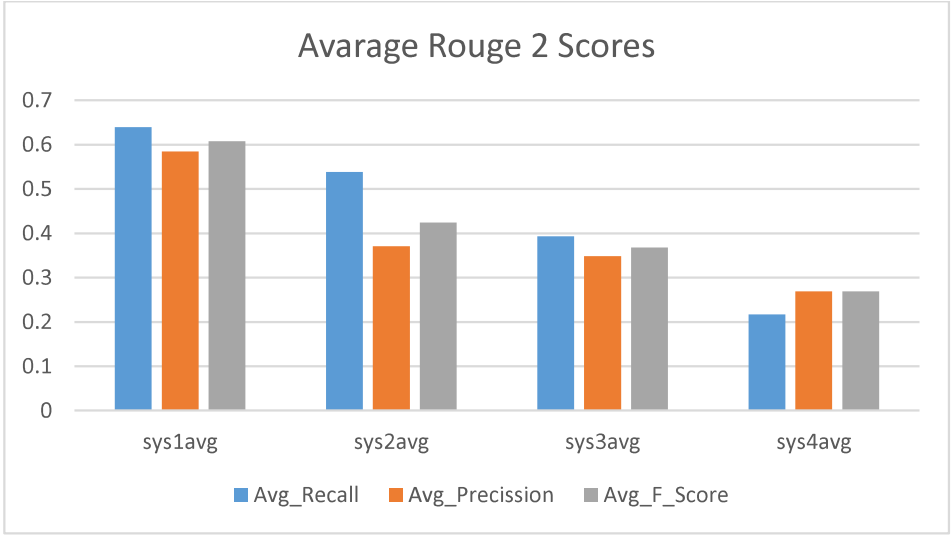


Figure 4.12. Chat comparison of average Rouge-2 scores

From the table and the chat comparison the rating of the P, R and F score indicated that the best summarizer from the four is the Simple summarizer for both the Rouge-1 and Rouge -2 calculation.

4.2.4. Comparison between the Rouge and Usability Study Evaluation

The thesis aimed to introduce a new task-based method for evaluating summaries, this was by the use of the normal real world users who interacted with the summaries produced and gave an evaluation by going through certain tasks and filling a ranking scale. This method is popular in the HCI field where websites and normal daily use software or hardware are rated. As a comparison to the evaluation, a popular method in the summarization and machine translation research is used which is called the Rouge score calculator. Differences in the rating were observed, however both Rouge and the users rated the best summarizer as the Simple Summarizer, differences came in the ranking of the other summarizers where LSA the second best in users was ranked as the third by the Rouge Scores. Sentiment Analyser rated as the third in the user study was rated last in the Rouge Calculation. While finally Gensim summarizer rated by Rouge as the second best was rated the least in the user study.

The difference in the ranking between the Rouge and the users is proposed here as a result of both evaluators. In the user based evaluation, the more participants found the better the results to be achieved however the scope and resources for this study was limited and only a few users were able to be tested, in future studies the more users evaluated the better the evaluation. From the Rouge side, the summary used as reference or gold standard method is seen as a bias due to the fact that humans differ in view point, multiple summaries can be produced by different users for the same article based on an angle or even on user might produce multiple summaries for the same article based emotion, experience or view point change. Agreeing on the top most important sentences is a difficult task on its own, hence this gives more importance to the proposed method for evaluation where the users based on their experience and how their needs are met by the summariser could form a better ranking system than the evaluators that depend part on human and part on machine, a fully user centred evaluation metrics as proposed in this thesis is termed as important.

5. RESULTS AND SUGGESTIONS

Text summarization is a technological art that aims at reducing the time and effort needed by the average user in consuming an understanding the information given. Daily news is an important factor in our current life where everyone wants to keep up with the ongoing real world events. Although a lot of sources give instant reports and updates and the internet is overflowed with information, not all the resources can be categorized as useful some of them are just filled with junk leaving the user with more information to go through and little output to show which is discouraging. This is where importance of technologies like text summarization come in. The thesis based on text summarization explored the journey that is Text Summarization giving all the background information and briefly detailing the interrelated fields that are interacted with in the process of designing the algorithms that carry out the task of summarization. Comparative study was done to test the real impact of text summarization for the average user in the real world, and the results were compared to the results produced by the standard industry evaluation Metric Rouge. Although the two evaluation metrics differed in the ranking order, the best summarizer was found to be the simple summarizer by both evaluation. Using the four summarizers, the Gensim TextRank-based, the Sumy LSA-based, the simple and the Sentiment Analysis-based summarizers, the hypothesis was investigated by answering the questions posed in the introduction part of the thesis. Even though the users favoured one summarizer over the other, the study supported the hypothesis that yes, using automatic summarizers is more efficient than blindly digging for information like the needle in the haystack parable. The two questions were answered in the affirmative where majority of the users supported the fact that a summarizer was able to identify read-worthy articles from discard-able ones and also from just the summaries the event behind the article and the important information could be gotten without the need for the article which is enough to prove that yes, regardless of the type and limitation of the summarizers, they are an effective way to keep up to date with the day to day real life happening which is all an average user could require. The usability scale borrowed from HCI also was found to be effective in measuring the efficiency of the summarizer,

using the feedback given, the calculation done proved the best summarizer in terms of usability was the simple summarizer. Moving forward this evaluation method is proposed as one of the task based evaluation method that us user centric.

Work in text summarization although advanced, it has yet to end or be termed as a solved problem. The future of text summarization lies in hybrids and deep learning techniques. Combination of different algorithms could be explored for extractive text summarization which could result in effective automatic summarization. While exploring capabilities in deep learning will ensure abstractive text summarization improves in terms of research.

In the evaluation sector although there is a positive aspect in automating evaluation process, the central user should not be neglected. In this research due to the limited nature of the scope and the resources needed, few users were tested for feedback, in future usability studies should aim at having a large group of users in order to give a more substantial results with no bias in any form. Users in future work could be grouped into different groups that entails ensuring each summarizer deals with each article and enough user to ensure no repeat reading is done. Work to develop this evaluation method will be proposed for the next step.

REFERENCES

[1] Can U., Alatas B., Big Social Network Data and Sustainable Economic Development, *Sustainability*, 2017, **9**(11), 2027.

[2] Luhn H.P., The Automatic Creation of Literature Abstracts, *IBM Journal of Research and Development,* 1958, **2**(2), 157-165.

[3] Omar N.A., Duru N., Text Summarization And Evaluation Methods:–An Overview, *Int. Journal of Engineering Research and Application*, 2017, **7**(12), 89-93.

[4] Human-Computer Interaction (HCI), https://www.interactiondesign.org/ literature/ topics/human-computer-interaction (Retrieved Date: 20 May 2018).

[5] McKeown K., Radev D., Generating Summaries of Multiple News Articles, *SIGIR Forum (ACM Special Interest Group on Information Retrieval),* New York, USA, 9-13 July 1995.

[6] Brandow R., Mitze K., Rau L., Automatic Condensation of Electronic Publications by Sentence Selection, I*nformation Processing & Management*, 1995, **31**(5), 675-685.

[7] Niu J., Zhao Q., Wang L., Chen H., Atiquzzaman M., Peng F., OnSeS: A Novel Online Short Text Summarization Based on BM25 and Neural Network, *IEEE Global Communications Conference (GLOBECOM),*Washington DC, USA,4-8 December 2016.

[8] Sah S., Kulhare S., Gray A., Venugopalan S., Prud'hommeaux E.T., Ptucha R.W, Semantic Text Summarization of Long Videos, *IEEE Winter Conference on Applications of Computer Vision (WACV),* Santa Rosa CA, 27-29 March 2017.

[9] Alami N., Meknassi M., Ouatik S.E., Ennahnahi N., Arabic text summarization based on graph theory, *IEEE/ACS 12th International Conference of Computer Systems and Applications* (*AICCSA),* Marrakech, Morocco, 17-20 November 2015.

[10] Radev D.R., Allison T., Blair-Goldensohn S., Blitzer J., Çelebi A., Dimitrov S., Drábek E., Hakim A., Lam W., Liu D., Otterbacher J., Qi H., Saggion H., Teufel S., Topper M., Winkel A., Zhang Z., MEAD - A Platform for Multidocument Multilingual Text Summarization, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 26-28 May 2004.

[11] Singh S.P., Kumar A., Mangal A., Singhal S., Bilingual automatic text summarization using unsupervised deep learning, *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, Tamilnadu, India, 3-5 March 2016.

[12] Marujo L., Ribeiro R., Gershman A., Matos D.M., Neto J.P., Carbonell J.G., Event-based summarization using a centrality-as-relevance model, *Knowledge and Information Systems,* 2016, **50**, 945-968.

[13] Daniel N., Radev D.R., Allison T., Sub-Event Based Multi-Document Summarization, *Proceedings of the HLT-NAACL 03 on Text summarization workshop* - Volume **5**, Edmonton, Canada, 27 May-1 June 2003.

[14] Lin C., Hovy E.H., the Automated Acquisition of Topic Signatures for Text Summarization, *18th International Conference on Computational Linguistics COLING*, Germany, 31 July- 4 August 2000.

[15] Wasson M., Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications, *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL),* Canada, 10-14 August 1998.

[16] Kupiec J., Pedersen J.O., Chen F., A Trainable Document Summarizer, *Proceedings of the 18th Annual International {ACM} {SIGIR} Conference on Research and Development in Information Retrieval,* Seattle, Washington, USA, 9-13 July 1995.

[17] Osborne M., Using Maximum Entropy for Sentence Extraction*, Proceedings of the ACL-02 Workshop on Automatic Summarization,* Philadelphia, Pennsylvania,11-12 July 2002.

[18] Neto J.L., Freitas A.A., Kaestner C.A., Automatic Text Summarization Using a Machine Learning Approach, *Advances in Artificial Intelligence, 16th Brazilian Symposium on Artificial Intelligence*, Porto de Galinhas/Recife, Brazil, 11-14 November 2002.

[19] Jo T., K nearest neighbour for text summarization using feature similarity. *International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE),* Khartoum, Sudan, 16-17 January 2017.

[20] Barzilay R., Elhadad M., Using Lexical Chains for Text Summarization, *In Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization,* Madrid Spain, 11 July 1997.

[21] Mitra M., Singhalz A., Buckleyyy C., Automatic Text Summarization by Paragraph Extraction, *In Proceedings of the Workshop on Intelligent Scalable Summarization at the ACL/EACL Conference*, Madrid, Spain, 11 July 1997.

[22] Jafari M., Wang J., Qin Y., Gheisari M., Shahabi A.S., Tao X., Automatic text summarization using fuzzy inference, *22nd International Conference on Automation and Computing (ICAC)*, Colchester, United Kingdom, 7-8 September, 2016.

[23] Wang S., Zhao X., Li B., Ge B., Tang D., Integrating Extractive and Abstractive Models for Long Text Summarization, *IEEE International Congress on Big Data (BigData Congress)*, Boston, MA, USA, 11-14 December 2017.

[24] Li A., Jiang T., Wang Q., Yu H., The Mixture of Textrank and Lexrank Techniques of Single Document Automatic Summarization Research in Tibetan, *2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 2016, **01**, 514-519.

[25] Dave H., Jaswal S., Multiple Text Document Summarization System using hybrid Summarization technique, *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, Dehradun, Uttrakhand, India,4-5 September 2015.

[26] Yadav N., Chatterjee N., Text Summarization Using Sentiment Analysis for DUC Data, *2016 International Conference on Information Technology (ICIT)*,Bhubaneswar, India, 22-24 December 2016.

[27] Edmundson H.P. New Methods in Automatic Extracting, *J. ACM*, **16**, 264-285, 1969.

[28] Mani I., Bloedorn E., Multi-document summarization by graph search and matching, *The Ninth Conference On Innovative Applications Of Artificial Intelligence (AAAI/IAAI)*, Providence, Rhode Island, 27-31 July 1997.

[29] Mani I., Gates B., Bloedorn E., Improving Summaries by Revising Them, *ACL. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL 99),* College Park, Maryland, 20-26 June 1999.

[30] Radev D.R., Otterbacher J., Winkel A., Blair-Goldensohn S., NewsInEssence: summarizing online news topics, *Commun. ACM*, 2005, **48**, 95-98.

[31] Conroy J.M., O'Leary D.P., Text Summarization via Hidden Markov Models, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* , New Orleans, Louisiana, USA, 19-12 September, 2001.

[32] Gong Y., Liu X., Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* , New Orleans, Louisiana, USA, 19-12 September, 2001.

[33] Mihalcea R., Tarau P., TextRank: Bringing Order into Text, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain 25-29 July 2004.

[34] Jagadeesh J., Pingali P., Varma V., Sentence extraction based single document summarization, *Workshop on Document Summarization,* Allahabad, India, 19-20 March 2005.

[35] Erkan G., Radev D.R., LexRank: Graph-based Lexical Centrality as Salience in Text Summarization, *J. Artif. Intell. Res.*, 2004, **22**, 457-479.

[36] Feldman R., Sanger J., *The Text Mining Handbook - Advanced Approaches in Analysing Unstructured Data*, 1st ed., United States of America, Cambridge University Press, 2007.

[37] Gharehchopogh F.S., Khalifelu Z.A., Analysis and evaluation of unstructured data: text mining versus natural language processing, 5th *International Conference on Application of Information and Communication Technologies AICT*, St. Maarten, The Netherlands, 20-25 March 2011.

[38] Kroeze J.H., Matthee M.C., Bothma T., Differentiating Between Data-Mining and Text-Mining Terminology, *South African Journal of Information Management,* 2004, **6**(4), 353-356.

[39] Solka J.L., Text Data Mining: Theory and Methods, *Statistics Surveys,* 2008, **2**, 94-112.

[40] Witten L. H., Don, Katherine J., Dewship, M., Tablan V., Text Mining in a Digital Library, *International Journal on Digital Libraries,* 2004, **4**(1), 56-59.

[41] Kao A., Poteet S.R., Natural Language Processing and Text Mining, *Springer Science+Business Media, LLC,* United States of America, 2007.

[42] Bista N., Best 3 Things To Learn About Data Mining vs Text Mining, EDUCBA, https://www.educba.com/data-mining-vs-text-mining/ (Retrieved Date: 30 March 2018).

[43] Kiser M., Introduction to Natural Language Processing (NLP), Retrieved from http://blog.algorithmia.com/introduction-natural-language-processing-nlp/ (Retrieved Date: 30 November 2017).

[44] Couto J., The Definitive Guide to Natural Language Processing, Retrieved from https://blog.monkeylearn.com/the-definitive-guide-to-natural-language-processing/ (Retrieved Date: 20 November 2017).

[45] Manaris B.Z., Natural Language Processing: A Human-Computer Interaction Perspective, *Advances in Computers*, 1998, **47**, 1-66.

[46] Herbrich R., Graepel T., *Handbook of Natural Language Processing*, 2nd ed., Chapman & Hall/CRC, United States of America, 2010.

[47] Calderon P., Bag of Words and Tf-idf Explained, Data Meets Media, http://datameetsmedia.com/bag-of-words-tf-idf-explained/ (25 September 2017).

[48] Machine learning (ML) (n.d), Retrieved From https://searchenterpriseai.techtarget.com/definition/machine-learning-ML (Retrieved date 05 June 2018).

[49] Alpaydin E., *Introduction to Machine Learning*, 2nd ed., The MIT Press, Cambridge, Massachusetts, London, England, 2010.

[50] Shwartz S.S., David S.B., *Understanding Machine Learning from Theory to Algorithms*, 1st ed., Cambridge University Press, United States of America, 2014.

[51] Heath N., What is machine learning? Everything you need to know, ZDNet, https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/ (Retrieved Date: 20 April 2018).

[52] Mitchell T.M., *Machine Learning*, McGraw-Hill Science/Engineering/Math, 1997.

[53] "Summary"dictionary.cambridge.org,https://dictionary.cambridge.org/dictionary/english/summary (Retrieved: 25 May 2018).

[54] "Summary",en.oxforddictionaries.com,https://en.oxforddictionaries.com/definition/summary, (Retrieved: 25 May 2018).

[55] "Summary",macmillandictionary.com,https://www.macmillandictionary.com/dictionary/british/summary_1 (Retrieved: 25 May 2018).

[56] "Summary",techopedia.com,https://www.techopedia.com/definition/25911/automatic-summarization (Retrieved: 25 May 2018).

[57] Reeve L.H., Han H., Nagori S.V., Yang J.C., Schwimmer T.A., Brooks A.D. Concept frequency distribution in biomedical text summarization, *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM)*, Arlington, VA, USA, 5-11 November 2006.

[58] Meena Y.K., Gopalani D., Domain Independent Framework for Automatic Text Summarization, *Procedia Computer Science*, 2015, **48**, 722-727.

[59] Shetty A., Bajaj R., Auto Text Summarization with Categorization and Sentiment Analysis, *International Journal of Computer Applications*, 2015, **130** (7), 57-60.

[60] Moratanch N., Chitrakala S., A survey on abstractive text summarization, *International Conference on Circuit, Power and Computing Technologies (ICCPCT),* Kumaracoil, India, 18-19 March, 2016.

[61] Siddharthan A., Tutorial on Abstractive Text Summarization, [power point slides] Retrieved from https://pdfs.semanticscholar.org/presentation/3d60/4c4c5acf3287870e7d478490aa0a26121396.pdf (Retrieved Date: 20 December 2017).

[62] Kasture N.R., Yargal N., Singh N.N., Kulkarni N.S., Mathur V.P., Mathur V.B., A Survey on Methods of Abstractive Text Summarization, *International Journal For Research In Emerging Science And Technology*, 2014, **1**(6), 53-57.

[63] Ferreira R., Freitas F.L., Cabral L.D., Lins R.D., Lima R., Silva G.D., Simske S.J., Favaro L., A Context Based Text Summarization System, *2014 11th IAPR International Workshop on Document Analysis Systems*, Tours – Loire Valley, France, 7-10 April 2014.

[64] Hogenboom F., Frasincar F., Kaymak U., Jong F.D., Caron E., A Survey of event extraction methods from text for decision support systems, *Decision Support Systems*, 2016, **85**, 12-22.

[65] Zhang Y., Szabo C., Sheng Q.Z., Improving Object and Event Monitoring on Twitter through Lexical Analysis and User Profiling, *Web Information Systems Engineering (WISE)*, Shanghai, China, 8-10 November 2016.

[66] Duke A., Robin Williams dead; family, friends and fans are 'totally devastated', CNN.com, http://edition.cnn.com/2014/08/11/showbiz/robin-williams-dead/index.html (Retrieved: 02 May 2018).

[67] Rehurek R., Gensim Topic Modelling for Humans https://radimrehurek.com/gensim/summarization/summariser.html, (Retrieved: 25 April 2018).

[68] Moore J., Uhuru Kenyatta Is Declared Winner of Kenya's Repeat Election, nytimes.com, https://www.nytimes.com/2017/10/30/world/africa/kenya-election-kenyatta-odinga.html (Retrieved: 02 May 2018).

[69] Sumy, https://pypi.org/project/sumy/, (Retrieved: 30 April 2018).

[70] Bhagwant, Latent Semantic Analysis (LSA) Tutorial, technowiki.wordpress, https://technowiki.wordpress.com/2011/08/27/latent-semantic-analysis-lsa-tutorial/ (Retrieved Date: 15 April 2018).

[71] Allahyari M., Pouriyeh S., Assefi M., Safaei S., Trippe E.D., Gutierrez J.B., Kochut K.J., Text Summarization Techniques : A Brief Survey, **arXiv:1707.02268**, (Retrieved: 30 December 2017).

[72] Steinberger J., Jezek K., Text summarization and singular value decomposition, *Advances in Information Systems Third International Conference (ADVIS),* Izmir, Turkey, 20-22 October, 2004.

[73] Murray G., Renals S., Carletta J., Extractive Summarization of Meeting Recordings, *Proceedings of the 9th European Conference on Speech Communication and Technology,* Lisbon, Portugal, 4-8 September 2005.

[74] Ozsoy M.G., Cicekli I., Alpaslan F.N., Text Summarization using Latent Semantic Analysis, *Journal of Information Science*,2011, **37**(4), 405-417.

[75] McKinley JR J.C., Bombings In East Africa: The Overview; Bombs Rip Apart 2 U.S. Embassies In Africa; Scores Killed; No Firm Motive Or Suspects,nytimes.com, https://www.nytimes.com/1998/08/08/world/bombings -east-africa-overview-bombs-rip-apart-2-us-embassies-africa-scores.html (Retrieved Date: 2 May 2018).

[76] Liu B., Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies*, 2012, **5**(1), 1-167.

[77] Brownlee J., How to Prepare Movie Review Data for Sentiment Analysis, machinelearningmastery.com, https://machinelearningmastery.com/prepare-movie-review-data-sentiment-analysis/ (Retrieved Date: 20 December 2018).

[78] Pang B., Lee L., A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts, *Association for Computational Linguistics (ACL)*, Barcelona, Spain, 21-26 July 2004.

[79] Green R.A., William and Catherine Marry in Royal Wedding at Westminster Abbey,CNN.com,http://edition.cnn.com/2011/WORLD/europe/04/29/uk.roya l .wedding.kate.william/index.html (Retrieved Date: 02 May 2018).

[80] Steinberger J., Jezek K., Evaluation Measures for Text Summarization. Computing and Informatics, 2009, **28**, 251-275.

[81] What is ROUGE and how it works for evaluation of summaries?, text-analytics101, http://text-analytics101.rxnlp.com/2017/01/how-rouge-works-for-evaluation-of.html (Retrieved Date: 20 May 2018).

[82] System Usability Scale (SUS), usability.gov, https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html (Retrieved Date: 25 April 2018).

[83] Thomas N., How to Use the System Usability Scale (SUS) To Evaluate Usability Your Website, http://usabilitygeek.com/how-to-use-the-system-usability-scale-sus-to-evaluate-the-usability-of-your-website/, (Retrieved Date: 30 April 2018).

APPENDIX

Appendix-A

Consent Form (Adult)

Before signing the form please note the following three points:

The information you give is intended to be used in a thesis study which targets to test how happy you are with the summarizer, it gives in a few line the summary of news articles.

The information you give will be used for the project only, it is private and will not be shared with third parties.

This is a volunteer study, it entails finishing up the tasks given and answering a questionnaire afterward, I therefore thank you for accepting to be part of this, however you are allowed to leave any time with no repercussions, completion of tasks is not compulsory. This is to test the summarizer and not you.

I agree to participate in the study conducted by Najma Omar to be used for thesis study.

I understand that participation in this usability study is voluntary and I agree to immediately raise any concerns or areas of discomfort during the session with the study administrator.

Please sign below to indicate that you have read and you understand the information on this form and that any questions you might have about the session have been answered.

Date: _____

Please write your name: _____

Please sign your name: _____

Thank you!

We appreciate your participation.

Appendix-B

Tasks: Activities for the Users to Test Usability of the Summarizer

Before beginning this activities please make sure to read and understand the given consent paper and sign it. You are encouraged to think aloud throughout the process, anything confusing or not in place is encouraged to be verbally stated, keep in mind that we are testing the system, not you. Help us evaluate the system better. We promise to protect your personal details and needed privacy.

For all the summaries given do the following:

Read the given article summary.

How long did it take you to read the summary?

What is the article about based on just the summary?

What is the event category based on the summary (events are happening i.e. Death, Terror etc.)

Does the summary give you enough idea on whether you want to read the full article for details?

Does the summary affect your attitude towards the whole article?

Read the source article.

How long did it take you to read the source article?

Did the article contain extra information to improve on your understanding of the context of the article?

Did reading the article after the summary, change your opinion on the story?

Fill the corresponding usability form from your experience with the summary

Thank you.

Appendix-C

System Usability Scale (provided in the CD folder) - Used for Ranking the summarizers via user survey.

Appendix-D

System Usability Scale Calculator (provided in the CD Folder) an excel worksheet used for calculating and interpreting the scores of the system usability scale.

PERSONAL PUBLICATIONS AND WORKS

[1] **Omar N.A**., Duru N., Text Summarization and Evaluation Methods:–An Overview, *Int. Journal of Engineering Research and Application*, 2017, **7**(12), 89-93.

BIOGRAPHY

Born 20 December 1989 in Mandera, Kenya. Completed Primary Education 2003 in Al-Huda Primary School, Kajiado, Kenya and Moi Girls Secondary School Isinya, Kenya in 2007. Studied Arabic Language and Undergraduate Degree BSc. Computer Science in International University of Africa, Khartoum, Sudan and Graduated in 2013. Worked in First Community Bank Kenya Computer Department 2014-15. 2015-2016 Joined Kocaeli University, Kocaeli, Turkey, 2016 Completed Certificate in Turkish Language. Currently (2018) a Master's Degree Student in Kocaeli University. Fluent in five Languages: Somali, Swahili, English, Arabic and Turkish.