| | |
|---|---|
| S | |
| E1 | |
| E2 | |
| **For Office Use Only** | |

# Masters Project Final Report

# (MCS)

# 2019

| | |
|---|---|
| **Project Title** | English news summarization from online sources |
| **Student Name** | M.Z.M Fiham |
| **Registration No. & Index No.** | 2016/MCS/030 - 16440302 |
| **Supervisor's Name** | Dr. A R Weerasinghe |

| **For Office Use ONLY** |
|---|
| |
| |

# English news summarization from online sources

**A dissertation submitted for the Degree of Master of Computer Science**

**M.Z.M Fiham**

**University of Colombo School of Computing**

**2019**

# Declaration

The thesis is my original work and has not been submitted previously for a degree at this or any other university/institute.

To the best of my knowledge it does not contain any material published or written by another person, except as acknowledged in the text.

Student Name: M.Z.M Fiham

Registration Number: 2016 / MCS / 030

Index Number: 16440302

_____

Signature: Date:

This is to certify that this thesis is based on the work of

Mr. M.Z.M Fiham

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Certified by:

Supervisor Name: Dr A R Weerasinghe

_____

Signature: Date:

# Acknowledgements

I am using this opportunity to express my gratefulness to everyone who supported me throughout the master's individual project. I am grateful for everyone's advice, guidance and constructive criticism for the project. I would like to thank my project supervisor Dr. A R Weerasinghe, a Senior Lecturer of University of Colombo School of Computing who has supported with valuable knowledge and vast experience.

# Abstract

Information are the most valuable assets in the current world. In the past people kept the information as physically. With the vast improvement of the information, they tried to store the information in digitally. People are reluctant to read large information. So, people tried to read the summary to understand without reading all the information. Summarization was done by the human because of no other mechanism. Manually summarization was so time consuming and high costly. Human resources for summarization also not available because of high demand and time consuming. Which ever the large content there is only a main idea that's included inside the content. Other information will be a descriptive information around the main idea. Motivation to find the main idea and show the summary to the readers.

Automatic text summarization introduces to the following contextual problem. This research is intended to find the English language news to be summarized with the most suitable approach and features to give high level of accuracy to the information. Using single document summarization with extractive methodology and abstractive techniques will be used in this research for generate the news summary.

By using above mentioned techniques achieved an English summary of the news article. With using abstractive technique leads the news article with simplified summary which can read and understand by any novice users. Above mentioned methodology will help to differentiate with the other researchers' outcomes.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 01 – Introduction

This chapter will explain the history of the text summarization and the how the different text summarization have evolved throughout the history to the present. This research is based on the English news summarization. Overview will discuss the history of the text summarization and their problems. Motivation will explain how authors interest to the following research. Aim and Objective discuss the intention and purpose to the research. Scope will explain what criteria the research will cover and what will not covered by this research.

## 1.1 Overview

News is an important message to the people. Day to day activities and updates can be getting to know by reading the news. There are multiple ways that news being delivered to the people. Printed and Digital news are the way of delivering the news. In our research project we are focusing on the Digital News. There are lot of different sources in the web that creates news for the readers. Most of the sources creates large size of the news articles for the readers. Current readers don't like to read lengthy content from the sources because they don't have time to read the all of it or they are not interested on the full article.

News author need to know about the readers before write the news article, because there will be experts and average persons will read on that area. Complicated and technical words will lead the user to skip the article or the source. For reduce the complexity of the content author uses describing words like adjectives. Because of using too many adjectives articles will be lengthy and readers will reluctant to read it.

Most web sources use manually summarize content to display the summary before on hand read the articles. This is to get the idea before reading it. Summary is a most important part when it come to a news media source. Most articles and breaking news are shown using a summary to readers. Currently summarizations are done by humans and same articles can be summarize in multiple way. Different authors summarization will not be unique for an article. Manual summarization will take more time and cost for the process. To do the summary manually we need

human involvement and we don't have lot of resources to do it. Because of this, Human considering doing the summarization automatically.

We are focusing on the domain of natural language processing to do the automatic summarization. Automatic summarization has two main approach, they are extractive and abstractive approach. In extractive approach it will use to select the sentences to generate the summary. Selection of the sentences is based on ranking the sentences which are appropriate to the given domain, article or title. In abstractive approach sentences are regenerated from the scratch to get the meaning of the article.

Goal of this automatic text summarization project is to be like human written summaries. We can understand the quality of the summarized content by answering the questions asked by source document with summarized content. If we can find the answers to who, what, when, where, why and how, summarized content will be likely to human written. Summarized content must be reducing the time of reading.

Motivation for this project is to resolve the automatic summarization with improve readability to the readers. Reduce the human involvement to do the text summarization. There have been researches going on based on text summarization and they have different approached to improve the readability and semantic relationship. Focus of this project to improve the readability and sematic relationship of the news summarized content. Most of the news which read by everyday have wordiness that makes the readers to be frustrated. Summarization using human involvement cost more and it's unbearable to news data sources. Summarization takes time and because of the competition with other news sources it's important to give a good summarization to keep the readers with the news source. This kind of facts motivated to work on this research topic.

Automatic summarization research has been started from 1958 by Luhun. He was started with the selection of the Top rank sentences for the summarization. Problem which still exist in current researches are they have not considered the sematic relationship of the sentences. Most difficult task of the summarization is to find the anaphora of the sentences before summarizing. Because of this entity recognition of the article will be mapping badly when summarizing the content. There are three different types of contents available to consider before doing the summarization. Structured, unstructured and semi structured. Most of the contents available on online are unstructured and semi structured. So, indexing on semi structured and unstructured data will be

difficult because of the descriptive of the content will be not so rich. When come to the summarization to find anaphoric relationship of the content will be so poor.

Named entity recognition is the other main important part of the problem which still recurring on current researches. When using the information extraction, we need to identify the word are belonging to which category and classify the words which are belongs to same category, we can use the categorization of the words when we are doing the stemming of the content to identify the duplicated declarations in the content. Same grouped words can be used for anaphora mapping in the sentences before do the summarization. Named entity recognition need large size of corpus to identify the it correctly. Training set of data with large number of corpus will reduce the error rate of the output.

Other problem which encountered in the current research is word sense disambiguation (WSD). This basically does is identifying the words meaning in the sentence. Because of the multiple meanings of the word, we need to identify the relevant meaning for the word inside the sentence to suit the consistency. Accuracy of current algorithm cannot be defining because building a corpus will be so expensive with human need to verify every word in the corpus. These are the current computer science problems encountered under this research topic.

## 1.2 Motivation

News is the way of getting the background information or incident to the user in lesser time limit. News authors are writing the news with their styles for exaggerate the news for getting the attention of the user. But the overall there is only a main idea that hidden in the content. So, we need to extract or recreate the content for gasp the idea. Motivation to extract the main idea from the content. User can get the decisions quickly based on the summarized content. They don’t need to waste the time for reading the whole content.

Whoever new to the English language cannot read and understand all the content, but he can understand the language simplified summarized content. Giving a solution for the problems will simplify the task to understand the information well.

## 1.3 Aims and Objective

Main aim of this research project is to generate the summary of news content and improve the sematic relationship of the content. The intention behind this is to reduce the human involvement in doing text summarization and aim to get a simpler and understandable content to summary.

For achieve this aims we need the following objectives.

- Named entity recognition and part-of-speech tagging will recognize the words categories and help to find correlation.
- Identify the correlation of the sentences before do the extraction of the content.
- To explain the content to the user, we need to use the simplified content, this will help the users who are new to the context and the language.
- Producing the syntaxial and semantical correct summary for the end user.

By achieving these objectives, research can produce a simplified summary to the readers without ambiguity sentences which lead the readers to decide.

## 1.4 Scope

In this project we are only focusing of the single document summarization. We only implement for the English news. We mainly focus on the extractive summarization, and we will apply some abstractive techniques (simplification) to extracted sentences. We use online sources to summarize the news. Treating all the news content and we don’t only limit for a domain or context.

# Chapter 02 – Background

There are lot of research works has been done for automatic text summarization. Most research has divided into based on extractive and abstractive summarization. Extractive is depending on selecting words and phrases from the original text. Abstractive summary will try to regenerate the whole meaning of the entire content with summarize sematic manner. Some researches focusing on human aided summarization, this is that main content will be highlighted to include in summary and human will be do the post processing of the summarization.

Furthermore, the extractive or abstractive summarization have divided into single source document summarization or multi-document summarization. Following Figure 1 will show the hierarchy of the summarization.

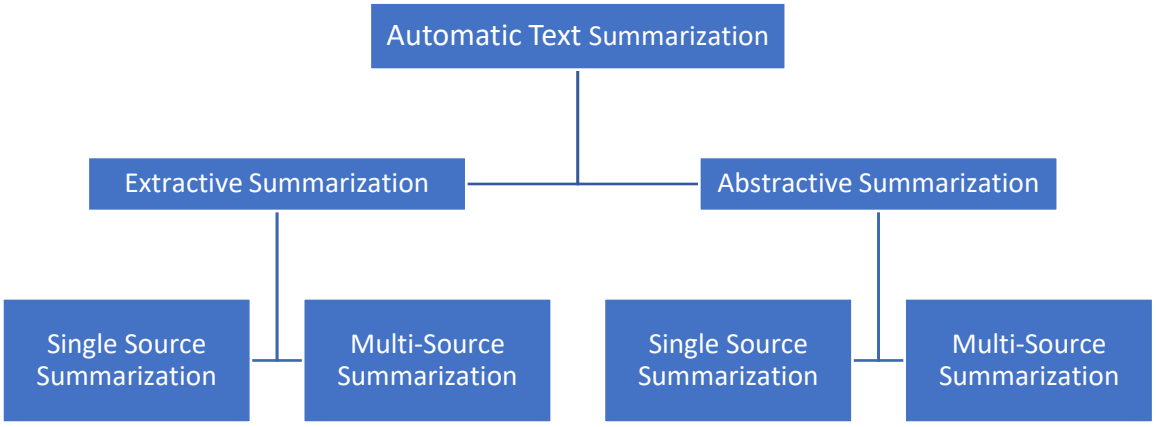


*Figure 1 Summarization Hierarchy*

Single source summarization will only get the content of its own and try to summarize content. Multi source summarization will get the articles which related to topic or the context and try to do the general summarization.

## 2.1 Extractive Summarization

In this approach only considering the extracted information from the original article and reorganize and display the relevant or selected sentences to the summary.

### 2.1.1 Single Source Summarization

In this section only explains the findings and problems faced by researchers which used extractive single source summarization.

History of the text summarization was started with H.P Luhn. He has started to find the automatic abstraction of the technical paper. He was focusing on the extractive single source summarization of the technical paper. He was influenced because of the need to eliminate the human effort to do the abstraction and he need to be generated bias abstraction. Because when come to human article summary will be influenced by his knowledge and context. And if he redoes the same summarization, he will come up with a difference set of summaries. To do the summarization he is doing the ranking of the sentences which is significant to the summary. To rank he need the significant factor, he uses the word frequency of the article and he find the sentences which are with the frequent words. He determines those sentences are the useful measurement for the summary.

With use of significance count of the words, Luhn used to identify the positions of words in a sentence are also significance for auto abstract summarization. He has eliminated the common words which used to tying the sentences. Because of the frequency of the tying words will be the highest when get the result and it will not help to get the knowledge from the document to extract the summary. He has found the tenses of the words will repeat in the article and need to consider those tense is as same word. Luhn's research has provided a simple way of extracting the subjective matter of the article and it's only giving an indicative abstract which only point out the subject key points. Using indicative summary, we can approximately generate the abstract of the summary. Luhn not discuss about the semantic relationship of the sentences which he is extracting from the technical article. There will be a problem that cannot tie the sentences because of the summary is generated from random sentences which are picked from the article. Next problem is some authors

will not use the same words in the documents and authors will express them with different angle of the sentences, which is not possible to identify with the Luhn research [1].

As the Researchers (Svore et al.,2007) they have used the single document summarization based on neural nets. They have not used any regeneration sentences from extracted contents. They have use the RankNet for rank the sentences. Basic steps of their research are to identify the similar sentences to the news three line of highlights to the extracted sentences and they will create a block summary from the sentences. Compare the extracted first line with first highlight and second with second and third with third. They will create a summary block from highlight also. They compare the two-summary block which was system generated. For extracting features from the sentences, they train a corpus and test with a test dataset. Then they apply to the real dataset for extraction [2]. Main problem from the following research depend on the human generated highlight sentences. There is no any limit for the highlights, and they take the minimum of three to compare. This researcher only focused on the extracting the information, but they have not treated semantic relationship of the sentences. RankNet requires a large size of labeled data to be trained before rank the real system and finding large numbers of labeled data will be so difficult. RankNet also have some problems, RankNet only work hard for ranking relevant document and less for the low relevant. They have used the ROUGE metric for evaluating the generated summary. ROUGE is a popular metrics for evaluating the automatic summarization. But there are lot of drawbacks with the metric. Metric cannot identify the synonyms of the words. Redundant information is ignored by the metric [3].

(Kupiec, Julian et al.,1995) they have proposed a statistical based framework for document summarizer. This research uses manual extraction to compare with the auto extract. He says document extraction can get the twenty percent information from the full text content. Author says combining all individual heuristic will give a good performance. They evaluate the extract with classification function success rate and the precision. But they need a labelled document training corpus for testing, but it's so expensive to find the dataset and they use a private data set for the training. They use Bayesian classification function for assign a score for the sentence and determine the sentences which include in the summary. They have faced with two main problems which they cannot isolate the title and the body. Second issue was presenting the sentences with scoring order will give a user unreadable format because of the missing of semantic relationship [6].

(Chin-Yew et al., 1997) they have been researching on a finding a best suite topic for the content or article. They have tested the possible and best places to find the topic from the original content. However, they are trying to extract a sentence for the topic. They are trying to develop a best position to find the topic with automatically. Some problems which the researchers ignored the morphological restructuring and anaphoric resolution. For ranking the sentences, they have depended on the keywords and the abstract of the document. They are ranking with high priority to the sentences which includes keywords or the abstract information. It's not suitable for the depending on those clues which given by the author of the document, because it will be misleading, or the research work will be narrowed to a specific situation [7].

(Yong, S. P et al.,2006) they have used a neural based text summarization system. They have used pre-processing for the content. They have done stop word removal and stemming. They have removed the suffix and prefix. They have used porter's algorithm for stemming [9]. But there are unavoidable errors which has exist in porter's algorithm. Over stemming error is one of the drawbacks. Under stemming errors are another kind of drawback that does not merge words which are to be connected [10].

Yong uses *tf-idf (w, s)* to find the significance factor of the sentences and words. Term frequency – inverse document frequency let you know how word is important for the document. Then they include the following numerical input to the neural network. From the output of neural network, they identify the special word included sentences. They redo the above-mentioned steps before output the summary. In their research they have not talked about any anaphora resolution or the sematic relationship of the sentences [10].

In 2000 Hongyan Jing was done a research on sentence reduction of automatic text summarization. They try to remove the unwanted sentences before doing the summarization. They first extract the sentences and do the reduction. As like the stemming, they do the reduction. To do the reduction they use syntactic knowledge, context and statistics computed. They have used five steps to do the reduction. Step 1 is the syntactic parsing; they use to generate a parser tree for the sentences in the document. Step 2 is the grammar checking; in this step they determine to keep the sentences which not appropriate for the reduction because of the grammar of the sentence will be out of order. They remove the prepositional phrase, adjectives ... etc. Step 3 is the context information; in this step they check the sentences with the topic, how related they are. They depend on the lexical links for find the relativeness. they check the morphological relation, find WordNet database with lexical

relations. They provide a score for sentences from the above methods and with the highest score sentences will be the most related. Step 4 is the Corpus evidence; they compare with a corpus that include human reduction sentences with original text to find the human removing format. They get the knowledge from the human practice with the corpus. Step 5 is the final decision; in this stage they will remove the sentences with above mentioned step results.

Evaluation Hongyan Jing compare the reduction sentences with the professional written reduction sentences. They have not found any solution for anaphoric resolution and sematic relationship of the sentences [11].

### 2.1.2 Multi Source Summarization

In this section only explains the findings and problems faced by researchers which used extractive multi source summarization.

(Haque, M etal.,2013) they have explained about the multi document summarization. And they have review different methods of doing the summarization. They tell that collecting the documents which related for the topics and extract the useful information from the models and extract the significance sentences from the useful information and reorder the sentences will give a human readable summary [4]. Edmundson in 1969 he was proposed a new method of automatic extracting the summary. He has tested the extraction with the four different parameters to check the accuracy, parameters are key words, cue words, title and heading words, structural indicators. Edmundson has used manually produced extract to compare with the system outputted extract. He expresses that in his research he has hot covered the mathematical symbols, citations foot notes and tables and figures. Edmundson says efficiency of the program need to be maximize because of the current is not enough. He also states that in future he needs to improve the sematic relationship of the content. Because of the extraction of the content redundancy will appear with the abstract summary. These are the problems reported with the Edmundson [5].

(Eduard et al.,1998) they have developed a summarization tool for do summary for the demand of the people. This framework is able of doing the multilingual text summarization. They are following three main techniques to generate the summary. First technique is identification of the relevant topic which related then they pass this output to the second technique interpretation they used to do the compression of the data, extraction of the sentences, generalize the sentences,

identification of the collected sentences. Final step is the generation of the summary; they use the interpretation of the multi-document to generate the summary. They have used Positioning technique to find the topic of the article in the first step. In topic interpretation they have used two concepts, first was generalization of the sentences which repeated all over the content and do the compression for content. They have created a signature for the word and categorized the signature and replaced with a target word. They have used a micro planner for the summary generation, but they have not succeeded for producing grammatical sentences. They have used two point for the evaluation of their output results. Compression ratio is a one of their evaluation which they tell summary need to be smaller than the original text. Then they tell it must only contain the information on the original text and no other new information which not include in the original text. They have specified the retention ratio with larger number will produce a good summary [8].

## 2.2 Abstractive Summarization

In this approach only considering original article and get the main idea from the article and regenerate the article summary with own words.

### 2.2.1 Single Source Summarization

In this section only explains the findings and problems faced by researchers which used abstractive single source summarization.

(Kavita et al.,2010) they have used a graph-based approach for single source summarization. Because of the difficulty of abstractive summarization, they have limited the work using prior knowledge and using neural language generation systems. With prior knowledge they use templates and frames to extract information. They use text regeneration from the deeper NLP analysis. In this approach researchers uses a graph to produce abstractive summary of highly redundant opinions. In this graph-based approach they encounter too much surface order of words. Because of that it cannot group the sentence at a deep semantic level[13].

(Paulus et al.,2017) they have used a neural network model and a novel intra attention. They provide a solution for RNN-based encoder-decoder models repeating phrase problem. They propose a new objective function to reduce exposure bias. They have only tested for the short inputs and outputs and they haven't used the long inputs and outputs[14].

### 2.2.2 Multi Source Summarization

In this section only explains the findings and problems faced by researchers which used abstractive multi source summarization.

(Liao et al.,2015) they have developed an abstractive multi-document summarization framework that can construct fine grained summary. They construct pool of concepts and facts. New sentences are generated selecting and merging informative phrases to maximize the salience of phrases. They use integer liner optimization for phrase selection and merge continuously to provide a better summary. There research is pending with quality of the grammar and the time optimizing for the framework to output the summary is pending[15].

(Lapalme et al.,2012) they have used a guided based summarization with multi document summarization to output hundred words. They have only limited to the domain specific summarization. Result of the research give a high density of information in the short summary. Because of the domain specific summarization, they only worked with a specific category. They have not tested in another category. They are going to expand the testing on other domains in future research[16].

# Chapter 03 – Methodology

News is great significance because of number of reasons for people. It will inform the people about events, notices, entertainment. News can be a media that will connect the people. People will get the idea what's happening around them.

Digital and traditional news are written by lengthy article. In the modern world people don't have time to read the lengthy article and people will be bored to read all. Recent research done by Microsoft has found that human attention is lesser than the goldfish. Goldfish can focus for nine seconds but human only can focus for 8 seconds. So, media writer tries to keep the focus of the reader when they are creating the content. Currently they give a summarization of the content that was written by the author. This summarization process will cost more time and money. Content writing media was worried to find writers to do summarization, because lack of written skill in the community and highly costed. So, they are requiring a solution by the current technology. We are proposing an automatic text summarization for the news content that will be easy to use and cost effective.

However, improvement of the current technology people starts to use smaller devices that can be suitable for their pockets. Because of the smaller screen size people doesn't read lengthy articles in smaller screens. So, users recommend the bit-sized content for the articles.

Before do the summarization we need to identify the topic of the content, interpreting and generate the summary for the content. We need to give a value for the sentences, words in the content to recognize the important content that can generate the main idea. Then we need to do the text simplification and coreference resolution. There we need to address the problem to remove technically and very often words to be identified and remap with the simple words to understand newbie users. Next problem is to resolve the anaphora where authors have used to refer the named entity with a pronoun or noun phrase that we need to identify what it is referring to.

Here we only consider the single document summarization. We use the extractive summarization with the abstractive technique. We use the online news sources to do summarization of the news and we don't consider of the domain of the content, we consider all contents are homogeneous. Some researchers are done based on the domain specific to generate more accurate summary, but

they are limited to the domain. Ex: Technical document summarization. So, this is not suitable for the generic problems.

First, we need to see the high-level model of the methodology. In our methodology we have decided to use the token simplification finally because we are removing most of the content and it will be a performance issue. Cataphora resolution is bit harder to implement because the pronouns that refer to the entity is occurring after the pronoun. Following is the high-level module of the system.

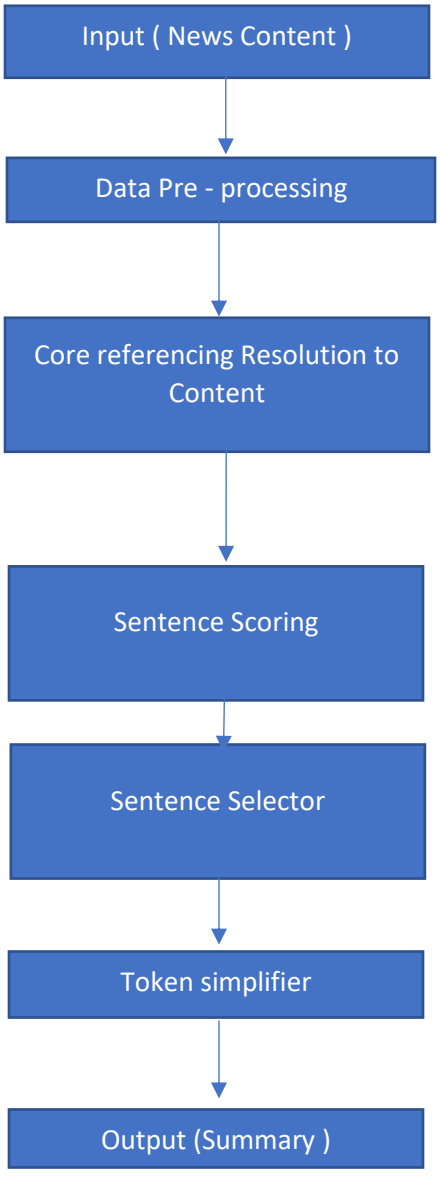


*Figure 2 Design High Level Model*

## 3.1 Input (News Content)

We are using an online news web application interface to get the English news. Upon getting the content from the application interface we will treat only the body of the news. We have used article scrapping library to extract the news from news url resources.

## 3.1 Data Pre-processing

For summary generation basic step, we need to follow was to pre-process the data. Necessary steps are the sentence boundary detection and tokenization. Commonly we are using part of speech tagging to identify the tokens based on context of the sentence. We are splitting the sentence from the content. Tokenize the sentence. We are using a sentence splitter to split the sentence, because it's not so easy to split the content with the different sentence stoppers. Specially tokenizer is the next step after splitting the sentence. We ignore the quoted sentences because that doesn't provide the content main points. We use the part of speech tagging to find the filter certain words like adjectives to ignore in the output.

## 3.2 Core Reference Resolution for Content

Most part of the researchers are doing the anaphoric resolution when doing the final selection of the sentence. But in the research identified that before do the data processing we can apply the anaphora resolve. Mostly the context is dependent on the expression's interpretations. We are focusing only the nominal expression dependencies on context. If the pronouns are preceding the entity this will called as cataphora. Finding a solution for dangling anaphora, we are going to use a coreference module for finding antecedent for anaphor.

We are using a third-party library to find the coreferences. Stanford CoreNLP Library. We use the coreference resolution system.

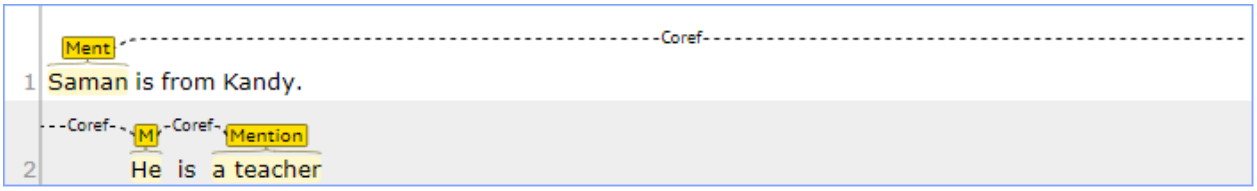


*Figure 3 Coreference Resolution (Ex.)* [12]

Here he is referred as Saman. Saman is the teacher. Library will provide the connected word references. We need to make the override information of the original content.

## 3.3 Sentence Scoring

In this section we are assigning a score to individual sentences that was splitter earlier. We are using a text rank-based approach to give a score to the sentences. For every sentence we use a vector representation (word embedding). Similarity between sentence vectors are calculate and create a matrix and store the value.

## 3.4 Sentence Selector

In here with using the matrix create a graph sentence as vertices and scores as edges to select the top custom value sentence to be in the summary. Output summary can see the top sentences are available. Here based on trial and error selected only four sentences to remain the meaning of the summary.

## 3.6 Token simplifier

In here we do replace the technical/hard words to the simplified words. Because novice users will be come to read the news which are not in their context. We need to cater the general crowd to be easy to understand the summary. In this approach we are using to find the synonyms from the WordNet and then we will feed to the Google Ngram viewer to find the most suitable word in the current context by the probability of the tokens. We use a library to find the probability based on the Google Ngram. Then we replace the main word of the content by the highest probability word.

This will be a good replacement of word because google Ngram has used the google books to check the word frequencies used by authors from past years and the current.

(Phrase) Ex:  digital asset

If we consider synonyms for the second gram of the words.

| **Synonym with Phrase** | **Frequency in Google NGram** |
|---|---|
| digital asset | 100% |
| digital property | 45% |
| digital goods | 100% |
| digital valuables | 0% |
| digital belongings | 0% |

*Table 1 Synonyms vs Frequency*

From here we select the highest frequency phrase. If the default word is 100% frequency, we use the phrase as it is.

## 3.7 Output (Summary)

Finally, there will be the most awaited solution that will be printed on the screen to view to the user.

By using this methodology, we will be delivering a summarized content to the novice user to the context or the domain. This will be a simplified version of the content that will be visualized to the user.

# Chapter 04 – Evaluation Plan

This chapter explains the evaluation plan for above methodology. There are multiple ways to evaluate this research project. In this research will select some of the evaluation plans. And explain the assumption and hypothesis and questions of this research.

Text summarization evaluation plan is the processes of rating the quality of the summary generated by the machine or human. There are many techniques that has been proposed and easiest and fastest method is need for the evaluation. From the evaluation best method of evaluation is human evaluation. But the human evaluation is costlier and time consuming. But human evaluation is more accurate. Automatic evaluation is faster and reusable. We must make the automatic translation is closer to the human translation. This will be a better output that matches in the evaluated plan. But the same content given to more than one human will be give different translation outputs. So, we cannot accept a more accurate translation. We need to focus to make the automatic translation to closer to the human translation.

Summary evaluation is a challenging task because of there is no standard or ideal summary for a document. Lack of standard evaluation metrics has also caused difficult to evaluate. To evaluate first, it needs to decide the important parts of the document is preserve with the output. This is a challenging task because of the identifying and displaying the output is the main task of the automatic summarization task which is there is not a standard methodology for the summarization task. When using the extractive summarization this task is bit easier compare to abstractive summarization. Abstractive summarization will be completely different from the original document. So, in abstractive case summarized content need to be compared with the original document they need to convert back to an intermediate representation. Then they need to check the readability of the summarization with compare to the original document. Grammatical evaluation of the output summary.

There are basically two type of evaluation, one is manually evaluation and other is automatically evaluation. Manually evaluation humans are involved. One way of doing the manually evaluation is for summarization is given original document and summarization check with the human score for the output. Other way is using the query-based summarization, where human can identify the how closest the answer which was given to the question. Human can give a score for the answer.

## 4.1 Questionnaire Based Evaluation

To Evaluate the extraction summary of the original content. Calculate score for raw extraction from the different steps of summarization. Its required to do number of steps to get a good summary from the original content. In this research hypothesis that only extraction of the sentence and representing the sentences doesn't give a meaningful summary. We need to consider other aspects like coreferencing, and the simplification of the words will give a good understand for the news readers. Summary of the content must answer the readers Five Ws questions. To achieve this summarization, This Research will do steps stated in the methodology. In each step we get raw data which we can compare with final summarization to check the accuracy and understandability of the summary.

To get a final score for the summary. In this research create an opinion and Interview based questionnaire to get a feedback from the user. With the Original content will produce the data cleaned version, coreference version of the summary. Sentence scored version. Simplified version. Template of the question are will be same and the jumble the versions of the answers in the template.

Each of the answer is the reduced version of the original content. For getting the response of the news article uses the google form. Can visualize the response of each users by graph and get the evaluation result. To evaluate following template will provided to user.

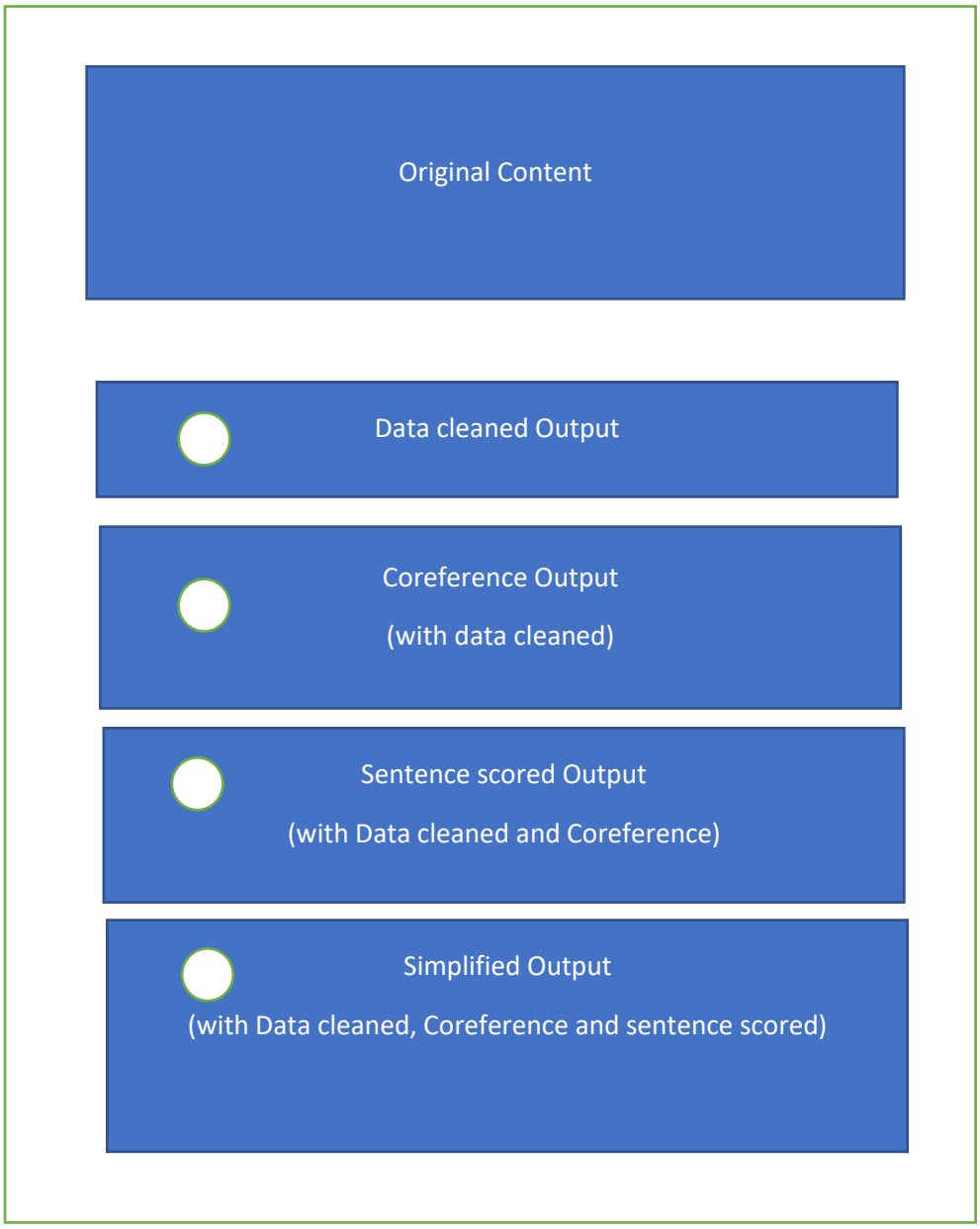


*Figure 4 Template for Evaluation*

Based on the questionnaire users will help to answer the correct and simplified and accurate summary for the relevant content. This questionnaire will give numerical scores for each option unbiasedly. With the numeric results can generate the graph and understand the hypothesis for this research. With the opinion following are the results which selected the simplified summary as a percentage.

| Article | Selected Data Cleaned Summary (%) | Selected Coreference Summary (%) | Selected Sentence Scored Summary (%) | Selected Simplified Summary (%) |
|---|---|---|---|---|
| A1 | 18.2 | 9.1 | 9.1 | **63.6** |
| A2 | 9.1 | 9.1 | 9.1 | **72.7** |
| A3 | 14.3 | **42.9** | 14.3 | 28.6 |
| A4 | 9.1 | 18.2 | 27.3 | **45.5** |
| A5 | 18.2 | 9.1 | 18.2 | **54.5** |
| A6 | 0 | 9.1 | 36.4 | **54.5** |

*Table 2 User opinion for the summary.*

A3 article has a less value to simplified summary is because it's meaning has lost when it's been sentence scored and simplified, In A3 article Coreference article has the higher percentage. Based on the percentages, Average high percentage has gone up for the simplified summary which make output of the methodology a good success rate. This give a positive feedback on methodology.

## 4.2 Readability Evaluation

We use another evaluation method called readability evaluation this is created for check readability test for English texts. We can get the understandability of the summarized text. Readability must be in a content to understand. Readability is one way of evaluating the quality of the content. In this research evaluating the previous six articles with readability index to identify the most readable content. For this evaluation uses the most popular readable metric called Flesh reading ease [17]. Highest score will indicate the content is easier to read and understand.

- Flesch reading ease score

$$206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) + 84.6\left(\frac{total\ syllables}{total\ words}\right) - 21.43$$

| Article | Original Article | Selected Data Cleaned Summary | Selected Coreference Summary | Selected Sentence Scored Summary | Selected Simplified Summary (Final Output) |
|---|---|---|---|---|---|
| A1 | 30.5 | **32.9** | 29.1 | 25.7 | 24.0 |
| A2 | 49.1 | **49.8** | 48.2 | 24.3 | 24.1 |
| A3 | 35.6 | **40.5** | 38.8 | 20.1 | 19.5 |
| A4 | 30.8 | **47.0** | 44.1 | 25.5 | 25.5 |
| A5 | 45.8 | **48.9** | 47.7 | 37.2 | 37.3 |
| A6 | 47.2 | **52.7** | 47.9 | 39.6 | 43.6 |

*Table 3 Readability Scores for the summary and Original Article.*

Based on the above readability scores for the original text and final output (Selected Simplified Summary) it has a small difference based on the score. Even the human written article doesn't score hundred, it has a value less than the fifty score for six articles. Averagely final output article has a better score with compare to the original article. In the summarization step different output level of article have more than better score than the original article. When selecting the sentences, the readability score is dropping because of the length of the content is decreasing and understandability is also decreasing. However, the average score of the final output simplified summary has a better score with compare to the original article. Final output summary can be understood and readable by collage level and above criteria peoples. Based on the values of readability score, the summary generated by the system can be understandable to the human.

# Chapter 05 – Conclusion and Future Works

This chapter will explain the output of the research project and achievements and future works to improve the existing problems.

## 5.1 Conclusion

In this research only focused on the English news summarization. There are vast number of researches done on English text summarization to identify a potential, efficient and accurate way to do the text summarization, but there is no better way found by the past researchers. Different researches have different problem faced on their research and they have kept the problems to future researches.

This research main focused to find a simple and accurate way to find the English text summary of the specific news article. So, this research has carried out past knowledge of text summarization and inbuilt libraries to get the output of the text summaries. Because this research not only focus on extractive summarization approach but also some techniques in abstractive approach. So, this research is based on hybrid approach to achieve a simplified summary. The result of the summary gives a successive summary which also can understand by novice users.

Based on the evaluation result cannot say the accuracy of the summary with numerical representation. But we can compare and see how good the summary. There are no hundred percent accurate evaluation. Researchers are building different metrics to evaluate the results. But lots of researchers are not compared with the human generated summary with the machine generated summary because it’s too hard to compare and even the different human or same human will summarize the same content in different ways.

Based on the evaluation results it shows the output on the mixed of coreference, sentence scored, selected and simplified summary is having a higher number of percentages which selected by the evaluating users. Evaluation of readability index provide the understandability of the content. This metrics score output provide eligible for college level and above people.

## 5.2 Future Works

This research was carried out with the knowledge that used in past researches. It is used because the prior knowledge has a specific outcome and can build the research on top of that to focus on a new problem without recurring the same work. Future work can be carried out for different direction. Current researches focusing on hybrid (extractive and abstractive) approach rather than selecting a single approach.

This research only focusses on the news article data set but it can focus on all other English content for summarization. Because this research can be used as general solution for summarize any content. There is a performance bottle neck of the summary model generator. Improvement for the performance to the summary generator framework need to be carried out in future. Simplifying the words used in output summary need to check all the words of the article with synonyms to check most used words. There is some situation which the words used for simplifying its always give the connectiveness or meaning for the sentence. So, identifying the specific words to replace in the summary based on the context of the sentence to be carried out in future work.

New evaluation method needs to build to find the accuracy of the summarization, because current evaluation is not strong enough to conclude the summary is perfectly correct. So future works are opened for evaluation of summary as well.

# References


[1] A. Creation and L. Abstracts, “Automatic Creation of Literature Abstracts*,” no. April, pp. 159–165, 1958.

[2] K. M. Svore, M. Way, L. Vanderwende, M. Way, C. J. C. Burges, and M. Way, “Enhancing Single-document Summarization by Combining RankNet and Third-party Sources - Svore et al. - 2007.pdf,” no. June, pp. 448–457, 2007.

[3] Ermakova, Liana, “Automatic summary evaluation. Roug e modifications,” 2012.

[4] M. Haque, M. M. Haque, S. Pervin, and Z. Begum, “Literature Review of Automatic Multiple Documents Text Summarization,” *Int. J. Innov. Appl. Stud. ISSN*, vol. 3, no. 1, pp. 2028–9324, 2013.

[5] H. P. Edmundson, “New Methods in Automatic Extracting,” *J. ACM*, vol. 16, no. 2, pp. 264–285, 1969.

[6] J. P. and F. C. Julian Kupiec, “A Trainable Document Summarizer,” *Proc. 18th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. retrieval. ACM*, 1995.

[7] C.-Y. Lin and E. Hovy, “Identifying topics by position,” *Proc. fifth Conf. Appl. Nat. Lang. Process.* -, pp. 283–290, 1997.

[8] E. Hovy and C.-Y. Lin, “Automated text summarization and the SUMMARIST system,” *Proc. a Work. held Balt. Maryl. Oct. 13-15, 1998* -, p. 197, 1996.

[9] S. P. Yong, A. I. Z. Abidin, and Y. Y. Chen, “A neural-based text summarization system,” *WIT Trans. Inf. Commun. Technol.*, vol. 37, pp. 185–192, 2006.

[10] R. S. Fadi Yamout, Rana Demachkieh, Ghalia Hamdan, “Further Enhancement to the Porter’s Stemming Algorithm,” pp. 7–23, 2004.

[11] H. Jing, “Sentence reduction for automatic text summarization,” pp. 310–315, 2000.

[12] “Stanford CoreNLP.” [Online]. Available: http://nlp.stanford.edu:8080/corenlp/. [Accessed: 24-Feb-2019].

[13] K. Ganesan, C. Zhai, and J. Han, “Opinosis : A Graph-Based Approach to Abstractive Summarization of Highly Redundant Opinions,” no. August, pp. 340–348, 2010.

[14] R. Paulus, C. Xiong, and R. Socher, “A DEEP REINFORCED MODEL,” no. i, pp. 1–12, 2017.

[15] L. Bing, P. Li, Y. Liao, and W. Lam, “Abstractive Multi-Document Summarization via Phrase Selection and Merging,” 2015.

[16] P. Genest and G. Lapalme, “Fully Abstractive Approach to Guided Summarization,” no. July, pp. 354–358, 2012.

[17] J. N. Farr, J. J. Jenkins, and D. G. Paterson, “Simplification of Flesch Reading Ease Formula.,” *J. Appl. Psychol.*, vol. 35, no. 5, pp. 333–337, 1951.

# Appendix A – Sample of Original Article, and Outputs in different step in summarization

## Original Article

As the White House mounted a furious assault on the Mueller report, its author and critics of a president not found to have conspired with Russia but not cleared of obstruction of justice, the chair of the House judiciary committee said obstruction, if proven, "would be [an] impeachable" offence.Trump and impeachment: where Democrats stand after Mueller Read more Trump's personal lawyer Rudy Giuliani appeared on multiple Sunday talk shows, arguing with and talking over interviewers in a series of chaotic encounters. On Fox News Sunday, he claimed Robert Mueller's 448-page report, which was released with redactions on Thursday, was full of "calumny, lies and distortion". On CNN's State of the Union, the former New York mayor went as far as to call one of Mueller's lawyers a "hitman" and claim the special counsel's team "came close to torturing people" in the questioning and confinement of Trump campaign chair Paul Manafort, who was convicted and sentenced on financial charges. Addressing the first volume of Mueller's report, which concerned Russian election interference and the Trump campaign's warm reception to "Russian offers of assistance", including an infamous June 2016 meeting with a Kremlin-linked lawyer offering "dirt" on Hillary Clinton, Giuliani told CNN: "There's nothing wrong with taking information from Russians. It depends on where it came from. "There's nothing wrong with taking information from Russians. It depends on where it came from Rudy Giuliani White House adviser Kellyanne Conway took a different tone, telling ABC's This Week: "The campaign that I managed in those last few months did not welcome help from Russia. In fact, I don't recall getting, being offered help from Russia. It would have been a ridiculous prospect. "In his second volume, Mueller considered potential obstruction of justice by Trump or his campaign, of which 11 possible instances were listed. He passed judgment on the issue to Congress. In opposition to Trump's blitz defense, House judiciary committee chair Jerrold Nadler told NBC's Meet the Press that if the evidence shows Trump obstructed justice, it would be an "impeachable" offence. "If proven, some of this would be impeachable, yes," Nadler said, adding that Democrats are not currently pursuing impeachment but plan to "go where the evidence leads". Democrats remain split on impeachment, which would begin in the House they control but almost certainly fail in the Republican-held Senate. Some fear it would galvanize his supporters and win him sympathy among floating voters. On Fox News Sunday, House intelligence chair Adam Schiff said to impeach or not to impeach was "going to be a very consequential decision and one I'm going to reserve judgment on until we have a chance to fully deliberate on it". House oversight chair Elijah Cummings told CBS's Face the Nation he could "foresee [impeachment] possibly coming". But he added: "I think we have to be very careful here. The American people, a lot of them clearly still don't believe that President Trump is doing things to destroy our democracy and has done a lot of things very poorly. "He also said he thought "history would smile upon us for standing up for the constitution". Giuliani expended significant energy attacking the credibility of former White House counsel Don McGann, a key witness cited by Mueller in descriptions of orders from Trump to fire the special counsel, an act McGann did not carry out-migrant's recollection of the order was "wrong", Giuliani said on CNN, claiming the experienced lawyer was "confused [and] cannot be relied upon". The Mueller

report depicts McGann taking notes of meetings with Trump, a practice Trump is said to have questioned. The Trump campaign has severed links with the law firm to which McGann returned. Facebook Twitter Pinterest Donald Trump and Melania Trump arrive at church for Easter services in Palm Beach, Florida. Photograph: Nicholas Kimm/AFP/Getty Images An incensed Giuliani went back and forth with CNN host Jake Tapper about the Mueller report. He made the "hitman" claim about Andrew Weissman, an experienced prosecutor Trump allies claim is too close to the Clintons. "I have no problem with investigating Russian interference in the election," Giuliani said, before downplaying possible foreign meddling. "The reality is, you think this is the first time the Russians have interfered with a presidential election. "Giuliani was pressed on criticism of Trump from the 2012 Republican nominee, Mitt Romney. The Utah senator said he was "sickened at the extent and pervasiveness of dishonesty and misdirection by individuals in the highest office of the land, including the president". Trump tampered with witnesses. These Senate Republicans voted to oust Bill Clinton for doing just that Read more "Stop the bull, stop this pious act," Giuliani said, adding: "There's nothing wrong with taking information from Russians. It depends on where it came from. "Former New York US attorney Preet Bharara, who was fired by Trump in 2017, told CNN: "That's an extraordinary statement and I would hope he would retract it." Giuliani ran against Romney for the 2008 Republican presidential nomination, both losing to John McCain. Asked if he would have accepted such information, Giuliani said: "I probably wouldn't. I wasn't asked. I would have advised, just out of excess of caution, don't do it." He also accused Romney of doing "things very similar" as a candidate, although he did not elaborate. Trump attacked Romney on Twitter on Saturday.Trump has repeatedly claimed Mueller's investigation fully exonerated him, which it did not, and called the inquiry a "hoax". He continued to tweet on Sunday, from his Mar-a-Lago resort in Florida.Attending church for Easter services, Trump was asked if he felt betrayed by staff members who spoke to Mueller. According to the White House pool report, the president "clearly heard the question" but "just smiled and turned away".

## Data Cleaned Article

As the White House mounted an assault on the Mueller report, its author and critics of a president not found to have conspired with Russia but not cleared of obstruction of justice, the chair of the House committee said obstruction, if proven, offence. On Fox News Sunday, he claimed Robert Mueller' s report, which was released with redactions on Thursday, was of. On CNN' s State of the Union, the New York mayor went as far as to call one of Mueller' s lawyers a and claim the counsel' s team in the questioning and confinement of Trump campaign chair Paul Manafort, who was convicted and sentenced on charges. Addressing the volume of Mueller' report, which concerned election interference and the Trump campaign' s reception to, including a June 2016 meeting with a lawyer offering on Hillary Clinton, Giuliani told CNN: There' s nothing with taking information from Russians. It depends on where it came from Rudy Giuliani White House adviser Kellyanne Conway took a tone, telling ABC' s This Week: In his volume, Mueller considered obstruction of justice by Trump or his campaign, of which 11 instances were listed. He passed judgment on the issue to Congress. In opposition to Trump' s blitz defense, House committee chair Jerrold Nadler told NBC' s Meet the Press that if the evidence shows Trump justice, it would be an offence. Nadler said, adding that Democrats are not currently pursuing impeachment but plan

to. Democrats remain split on impeachment, which would begin in the House they control but almost certainly fail in the Republican-held Senate. Some fear it would galvanise his supporters and win him among floating voters. On Fox News Sunday, House intelligence chair Adam Schiff said to impeach or not to impeach was. House oversight chair Elijah Cummings told CBS' s Face the Nation he could. But he added: He also said he thought. Giuliani expended energy attacking the credibility of White House counsel Don McGahn, a witness cited by Mueller in descriptions of orders from Trump to fire the counsel, an act McGahn did not carry out. McGahn' recollection of the order was, Giuliani said on CNN, claiming the lawyer was. The Mueller report depicts McGahn taking notes of meetings with Trump, a practice Trump is said to have questioned. The Trump campaign has severed links with the law firm to which McGahn returned. Facebook Twitter Pinterest Donald Trump and Melania Trump arrive at church for Easter services in Palm Beach, Florida. Photograph: Nicholas Kamm/AFP/Getty Images An Giuliani went back and forth with CNN host Jake Tapper about the Mueller report. He made the claim about Andrew Weissmann, a prosecutor Trump allies claim is too to the Clintons. Giuliani said, before downplaying meddling. Giuliani was pressed on criticism of Trump from the 2012 Republican nominee, Mitt Romney. The Utah senator said he was. Trump tampered with witnesses. Asked if he would have accepted information, Giuliani said: He also accused Romney of doing as a candidate, although he did not elaborate. Trump attacked Romney on Twitter on Saturday. Trump has repeatedly claimed Mueller' s investigation fully exonerated him, which it did not, and called the inquiry a. He continued to tweet on Sunday, from his resort in Florida. Attending church for Easter services, Trump was asked if he felt betrayed by staff members who spoke to Mueller. According to the White House pool report.

## Coreference Article

As the White House mounted an assault on the Mueller report, the White House's author and critics of a president not found to have conspired with Russia but not cleared of obstruction of justice, the chair of the House committee said obstruction, if proven, offence. On Fox News Sunday, he claimed Robert Mueller ' s report, which was released with redactions on Thursday, was of. On CNN ' s State of the Union, the New York mayor went as far as to call one of Mueller ' s lawyers a and claim the counsel ' s team in the questioning and confinement of Trump campaign chair Paul Manafort, who was convicted and sentenced on charges. Addressing the volume of Mueller ' report, which concerned election interference and the Trump campaign ' s reception to, including a June 2016 meeting with a lawyer offering on Hillary Clinton, Giuliani told CNN: There ' s nothing with taking information from Russians. It depends on where It came from Rudy Giuliani White House adviser Kellyanne Conway took a tone, telling ABC ' s This Week: In his's volume, Mueller considered obstruction of justice by Trump or his's campaign, of which 11 instances were listed. his passed judgment on the issue to Congress. In opposition to Trump ' s blitz defense, House committee chair Jerrold Nadler told NBC ' s Meet the Press that if the evidence shows Trump justice, the evidence would be an offence. Nadler said, adding that Democrats are not currently pursuing impeachment but plan to. Democrats remain split on impeachment, which would begin in the House Democrats control but almost certainly fail in the Republican-held Senate. Some fear it would galvanise his's supporters and win his among floating voters. On Fox News Sunday, House

intelligence chair Adam Schiff said to impeach or not to impeach was. House oversight chair Elijah Cummings told CBS ' s Face the Nation House oversight chair Elijah Cummings could. But House oversight chair Elijah Cummings added: House oversight chair Elijah Cummings also said House oversight chair Elijah Cummings thought. Giuliani expended energy attacking the credibility of White House counsel Don McGahn, a witness cited by Mueller in descriptions of orders from Trump to fire the counsel, an act McGahn did not carry out. McGahn ' recollection of the order was, Giuliani said on CNN, claiming the lawyer was. The Mueller report depicts McGahn taking notes of meetings with Trump, a practice Trump is said to have questioned. The Trump campaign has severed links with the law firm to which McGahn returned. Facebook Twitter Pinterest Donald Trump and Melania Trump arrive at church for Easter services in Palm Beach, Florida. Photograph: Nicholas Kamm/AFP/Getty Images a Giuliani went back and forth with CNN host Jake Tapper about the Mueller report. He made the claim about Andrew Weissmann, a prosecutor Trump allies claim is too to the Clintons. Giuliani said, before downplaying meddling. Giuliani was pressed on criticism of Trump from the 2012 Republican nominee, Mitt Romney. The Utah senator said The Utah senator was. Trump tampered with witnesses. Asked if The Utah senator would have accepted information, Giuliani said: The Utah senator also accused Romney of doing as a candidate, although The Utah senator did not elaborate. Trump attacked Romney on Twitter on Saturday. Trump has repeatedly claimed Mueller ' s investigation fully exonerated The Utah senator, which Mueller ' s investigation did not, and called the inquiry a. Trump continued to tweet on Sunday, from Trump's resort in Florida. Attending church for Easter services, Trump was asked if Trump felt betrayed by staff members who spoke to Mueller. According to the White House pool report.

## Sentence Scored Article

As the White House mounted an assault on the Mueller report, the White House's author and critics of a president not found to have conspired with Russia but not cleared of obstruction of justice, the chair of the House committee said obstruction, if proven, offence. Addressing the volume of Mueller ' report, which concerned election interference and the Trump campaign ' s reception to, including a June 2016 meeting with a lawyer offering on Hillary Clinton, Giuliani told CNN: There ' s nothing with taking information from Russians. It depends on where It came from Rudy Giuliani White House adviser Kellyanne Conway took a tone, telling ABC ' s This Week: In his volume, Mueller considered obstruction of justice by Trump or his campaign, of which 11 instances were listed. According to the White House pool report.

## Simplified Article

As the White House mounted an attack on the Mueller report, the White House 's author and critics of a President not found to have conspired with Russia but not cleared of obstruction of justice, the President of the House Committee said obstruction, if proven, crime. Addressing the volume of Mueller ' report, which concerned election noise and the Trump campaign ' reception to, including a June 2016 meet with a lawyer offer on Hillary Clinton, Giuliani told CNN: There ' s nothing with taking data from Russians. It depends on where It came from Rudy Giuliani White House adviser Kellyanne Conway took a spirit, telling ABC ' s This week: In his book, Mueller considered obstruction of judge by Trump or his campaign, of which 11 instances were listed. According to the White House pool report.