



University of Eastern Finland
School of Computing
Master's Thesis

AUTOMATIC KEYWORD EXTRACTION METHOD FOR MULTILINGUAL WEB PAGES

Shariful Islam Majumdar

July 2021

ABSTRACT

Finding desired resources is getting complex day by day due to the heavy amount of online data. To extend with, the amount of online data is being increased dramatically, more than 1.7 billion websites, so that users could not find the expected result easily. For instance, web documents could be classified by getting knowledge from the keywords which may give a concrete idea throughout the document immediately. Therefore, it is necessary to have an explicit keyword extraction method which will help search engines to retrieve the accurate data. To get over from this problem, we propose a new solution for automatic keyword extraction based on language and domain independent web pages. The idea relies on DOM structure and language detection where different DOM features (title tag, anchor tag, headers – h1 to h3, term frequency and URL – host and path) have been focused. After extracting the features by segmenting DOM structures, the candidate keywords are ranked based on the different positions of keywords. Then top ten (10) ranked keywords are considered as the extracted final keywords. However, the proposed method outperforms the other relevant methods like *TextRank* and *D-rank* for multilingual web pages.

Keywords: Keyword extraction, Language independent, DOM tree, Web page, Information retrieval

ACKNOWLEDGEMENT

First, I would like to express my heartiest thanks and gratefulness to almighty Allah for His divine blessing makes me possible to complete this research successfully.

Besides, I feel grateful to and wish my profound and indebtedness to **Professor Pasi Fränti**, University of Eastern Finland, Joensuu. Deep Knowledge and keen interest of my supervisor in the field of machine learning influenced me to carry out this research. His scholarly guidance, constructive criticism, valuable advice, reading inferior drafts and correcting them at all stages have made it possible to complete this thesis.

Moreover, I would like to express my heartiest gratitude to Himat Shah, for his kind help, continual encouragement, constant and energetic supervision to finish my project, and our coordinator Oili Kohonen for her continual and unconditional supports.

Thereafter, I would like to thank my entire course mates in University of Eastern Finland, who took part in this discuss while completing the course work.

Finally, I must acknowledge with due respect the constant support and patience of our parents as well as my beloved wife, Eshia Farnaj.

TABLE OF CONTENTS

Contents	Page
ABSTRACT	i
ACKNOWLEDGEMENTS	ii
CHAPTER 1. INTRODUCTION	1-5
1.1 Problem Definition	2-3
1.2 Motivation	4
1.3 Research Aim and Objectives	4
1.4 Contribution	5
1.5 Research Structure	5
CHAPTER 2. METHODS OF KEYWORD EXTRACTION	6-14
2.1 Statistical (Frequencies of Words)	6-9
2.2 Linguistic (POS Tagging and POS Patterns)	9-12
2.3 Structural (HTML and DOM Features)	13-14
CHAPTER 3. PROPOSED MULTILINGUAL KEYWORD EXTRACTION METHOD	15-24
3.1 New Approach for Extracting Keyword	15
3.2 DOM Structure	16
3.3 Feature Extraction	17-19
3.4 Text Extraction	19-20
3.5 Language Detection	20-22
3.6 Stop Words Removal	22-23
3.7 Keyword Selection	23
3.8 Evaluation Measures	23-24
3.9 Workflow	24
CHAPTER 4. EXPERIMENT	25-37

4.1 Data Collection	28-29
4.2 An Overview of Keyword Extraction Process	29
4.3 Keyword Extraction Steps	30-36
4.4 Technical Tools	37
CHAPTER 5. RESULT DISCUSSION	38-40
CHAPTER 6. CONCLUSION	41
REFERENCES	42-45

LIST OF ABBREVIATIONS

DOM	Document Object Module
POS	Part-of-Speech
WWW	World Wide Web
URL	Universal Resource Locator
CSS	Cascading Style Sheets
HTML	Hyper Text Markup Language
XML	Extensible Markup Language
XPATH	XML Path Language
NLP	Natural Language Processing
SVM	Support Vector Machine
IMDB	Internet Movie Database
RAKE	Rapid Automatic Keyword Extraction
IEEE	Institute of Electrical and Electronics Engineers
H1	Heading Size 1
H2	Heading Size 2
H3	Heading Size 3
a	Anchor
tr	Table Row
td	Table Data
http	Hypertext Transfer Protocol
href	Hypertext REference
RBF	Radial Basis Function

CHAPTER 1. INTRODUCTION

It is going to be difficult to manage data day by day due to the fast-growing online resources for the wonder of technological revolution. Extracting proper keywords could help to handle data in better way for instance, to categorize the articles in respective disciplines. On the other hand, by viewing the keywords of an article, readers could have a quick insight into the content to get the idea how relevant it is. This is because the keywords are the most significant and unique words which would be the most obvious description through the documents. Therefore, it is necessary to extract proper keywords for all kind of resources especially for the multilingual web pages. As we are aware that plenty of research have been done on this topic but still there are some scopes to study more on it as for multilingual pages. The existing methods, however, are not suitable enough to extract accurate keywords due to the use of more than one language in a single document.

In this research study, we focused on the term keyword which is one of the very common topics in web data mining especially in search engine optimization process to manage web data. *Keyword* is a word or an aspect of the subject which describe the document precisely [1]. According to International Encyclopedia of Information and Library Science [2], a keyword as *a word that succinctly and accurately describes the subject, or an aspect of the subject, discussed in a document*. On the other hand, *Keyword extraction* is a way of automatically identification of a collection of words or phrases which represent the whole document wisely [3]. Nowadays, keyword extraction is one of the key parts of web data mining which has wide range of uses to manage data like data searching, indexing, clustering, information processing, advertising and, so on.

Due to the increasing amount of online data rapidly, it is also necessary to retrieve original data to the user by extracting concise data from the heavy amount of similar junk data. Therefore, there are so many applications of keyword extraction especially for managing online data for instance, social networks, and location-based application [4]. It has also wide range of uses in different areas like searching web data, indexing, summarizing, highlighting, browsing, labelling

[5], clustering [6], industry informatics, classification [7], information retrieval, topic modeling [8], and content-targeting advertising [1, 9].

Moreover, the role of an application for keyword extraction is to define a set of words that best describe the document automatically in a text. These keywords can be useful entries to create an automated index for a group of documents.

1.1 Problem Statement

For the rapid growing of online data, it might be challenging task to find the expected data. Though search engines manage data to find it easier but still the performance of retrieving accurate data depends on the quality of keywords [10].

In our study, most of the websites are being well structured by meeting standard criteria, but still we can see many websites are formed in typical ways which need to be processed towards getting the proper keywords. In Figure 1, as we can see that it is formed with mixed mode languages particularly *English* and *Bengali* for which we need some sort of multilingual solution to extract potential keywords.

As we know that there are so many studies have been done on keyword extraction but most of them are depending on structural documents, and single language-based web page. Therefore, multilingual web pages need to be concentrated scholarly, where handling the more than one language would be the main challenging job. Moreover, web document does not follow any standard format so that we cannot depend on the body or any specific location to get the important information.

On the other hand, most of the solution focus on organized web pages like news, social, academics and so on but service-based web does not follow any standard format which crates problem to find the actual content because of the advertisements and scattered organization including unnecessary contents.

By contrast, web document is little bit different than text document. To explain with, web text is unstructured and heterogenous including irrelevant text (structure, tags, styles, hyper link, navigation menus, scripts codes, formatting, adds, and so on). On the other hand, document text

is well organized, standard writing format and rules, and homogenous in nature which is comparatively easier to handle.

The image shows the homepage of Bangladesh Open University. At the top, there is a search bar and the university's logo and name in Bengali and English: "বাংলাদেশ উন্মুক্ত বিশ্ববিদ্যালয় BANGLADESH OPEN UNIVERSITY (We assure education at your doorstep)". Below the header is a navigation menu with links: Home, Authority, Results, Academic Info, News/Notice, Schools, Divisions, RCs/SRCs, Webmail, Contact, and Publications.

The main banner features a photograph of several people, including officials and women in traditional attire, standing in front of a large stone relief sculpture. A circular floral wreath is placed in front of them. Below the photo, text in Bengali reads: "ঐতিহাসিক ৭ই মার্চ এ স্বাধীনতা চিরন্তন স্বারক ভাঙ্কর্থে জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমান-এর প্রতিকৃতিতে পুষ্পমালা অর্পণ" (Floral wreath offered to the portrait of the Father of the Nation, Sheikh Mujibur Rahman, on the historic 7th March, the eternal Swaraj Bhankor).

Below the banner is a pink header with the text: "বাং গাজীপুর ক্যাম্পাসে জাতির পিতা বঙ্গবন্ধু শেখ মুজিবুর রহমান – এর জন্মশতবার্ষিকী উদ্‌যাপিত হয়।** **জন্মশতবার্ষিকীতে জাতির পিতার প্রতি বিনয় শ্রদ্ধা ** ১৬ মার্চ" (The 100th birth anniversary of the Father of the Nation, Sheikh Mujibur Rahman, is celebrated in the Bangladesh Open University campus. ** Tribute and respect to the Father of the Nation on his 100th birth anniversary ** 16 March).

The main content area is divided into several sections:

- Mujib Corner:** A section dedicated to the 100th birth anniversary of Sheikh Mujibur Rahman. It features a portrait of him and the text: "মুজিববর্ষ" (Mujib Year), "জন্মশতবার্ষিকীতে জাতির পিতার প্রতি বিনয় শ্রদ্ধা" (Tribute and respect to the Father of the Nation on his 100th birth anniversary), and a list of activities: "বঙ্গবন্ধুর জন্মবার্ষিকী ২০২১ অনুষ্ঠানের ফটো গ্যালারী" (Photo gallery of the 2021 birth anniversary), "বঙ্গবন্ধুর জন্মশতবার্ষিকী অনুষ্ঠানের ফটো গ্যালারী" (Photo gallery of the 100th birth anniversary), "বঙ্গবন্ধুর জন্মশতবার্ষিকী অনুষ্ঠানের ভিডিও চিত্র" (Video of the 100th birth anniversary), "মুজিববর্ষের ক্ষণগণনা অনুষ্ঠানের ফটো গ্যালারী" (Photo gallery of the birth anniversary), "মুজিববর্ষের ক্ষণগণনা অনুষ্ঠানের ভিডিও চিত্র" (Video of the birth anniversary), "মুজিববর্ষের ক্ষণগণনা অনুষ্ঠানটি" (The birth anniversary), and "বাঙালির জাতিসত্তা ও বঙ্গবন্ধু – কবি ড. কামাল আবদুল নাসের চৌধুরী" (Bangladeshi Nationality and Sheikh Mujibur Rahman – Dr. Kamal Abdul Nasir Chowdhury).
- Mujib Karna Mujib Corner:** A section titled "Father of the Nation Bangabandhu Sheikh Mujibur Rahman" with a portrait and text: "বঙ্গবন্ধুর ফটো গ্যালারী" (Photo gallery), "ঐতিহাসিক ৭ই মার্চ এর ভাষণ" (Speech on 7th March), "ঐতিহাসিক ১০ই জুনুয়ারি এর ভাষণ" (Speech on 10th June), "মুক্তিযুদ্ধাদের অঙ্গসমর্পণ অনুষ্ঠানের ভাষণ" (Speech at the surrender ceremony), "প্রধানমন্ত্রী হিসাবে শপথ গ্রহণ ও পরবর্তী সংবাদ সম্মেলন" (Swearing in as Prime Minister and subsequent press conference), "জাতিসংঘে বাংলাদেশ ১ম ভাষণ" (1st speech in the UN), "সেনাবাহিনী কর্মকর্তাদের উদ্দেশ্যে ভাষণ" (Speech to the army officers), and "১৯৭০ সালে নির্বাচনের প্রেক্ষাগলে জাতির উদ্দেশ্যে বক্তার ভাষণ" (Speech at the election stage in 1970).
- করোনায় করণীয়:** A section titled "Care of Social Distancing and Hygiene".
- Institutional Info:** A section with links for "Prospective Students", "All Academic Programs", "Programs and Students", and "Functions of Divisions".
- About Bangladesh Open University:** A section with the text: "The need for an open university in Bangladesh was felt long ago. The history of distance education in Bangladesh dates back to 1956 when the..."

Figure 1. A Multilingual Web Page

1.2 Motivation

As per our study, the existing solutions are depended on language specific so there are still some scopes to research more for the multilingual based web documents. Moreover, our study covers the wide range of web pages including service-based web documents.

In perspective of solution, previously many researchers have proved [4] that title tag is the best sources to get most important information over any web documents which is considered for the current research as well with giving priority.

On the other hand, no language model has been applied beforehand which could play for multilingual web pages. Most of the solutions depend on the Natural Language Processing (NLP) complex structure to process the language, which is language specific, and does not support multiple languages at a time. Therefore, we looked forward for something which could handle multiple languages at a time. This is because of our targeted web pages would be multilingual.

To get over from any unwanted interruption we focus on own created dataset rather depending on any external sources, training data or supervision. During preparing dataset, we would look for multilingual websites as well as make sure that web pages have more than one keywords or key phrases in the meta tag with more than one used language. There are some external open sources applications for the analyzing meta tags which could make easier us to create a corpus.

1.3 Research Aim and Objectives

The main goal of this research is to handle multilingual web pages to extract keywords. Moreover, this research will deliver a complete solution which would avoid the machine learning strategy (supervised and unsupervised) as well as complex Natural Language Processing (NLP) structure. Besides, we will search for such an application which could manage more than one language at a time to detect the used languages over the web documents.

To extract multilingual keywords as an experiment, we would use our own dataset which would have multiple meta keywords including multiple languages so that we do not have to depend on any external resources.

1.4 Contribution

Beside the concentrating on keyword extraction method, we have experimented the new method to evaluate for multilingual web documents. Below is the list of key notes from our contribution during this research:

- Based on the DOM structure we have introduced a new method which will allow you to extract keywords from multilingual websites.
- We have created a dataset which is little bit different than others. In the dataset, we make sure that each web page has more than one meta keyword or key phrases as well as more than one language applied.
- The solution does not depend on any specific language.
- The proposed idea does not rely on any external sources.
- This search does not depend on machine learning or complex NLP structure.

1.5 Thesis Structure

This research article has been organized as follows: The relevant methods are discussed in the Chapter 2. The new approach of extracting keywords for multilingual web pages has been presented in Chapter 3. In Chapter 4 and 5, we have demonstrated our experiment including discussion of result comparisons and Chapter 6 covers the summary and future work of our study.

CHAPTER 2. METHODS OF KEYWORD EXTRACTION

Keywords are the most obvious description through a document. There are numerous studies have done on information extraction by concentrating different aspects. However, the focus of this study is to understand the key concepts regarding keyword extraction throughout a document especially for web documents. This literature review work has been done to be aware about the existing solutions as well as the overall impression of keywords in web pages.

2.1 Statistical (Frequencies of Words)

R. Mihalcea et al., 2004, presented a graph-based ranking model for text preprocessing [3] and demonstrates how this model can be applied in Natural Language Processing (NLP), which is known as TextRank. This study illustrates an unsupervised approach for extracting keywords and sentences. The methodology of this study has been classified into three main segments: for instance, *Text Ranking Model* (TRM), *Keyword Extraction* (KE) and *Sentence Extraction* (SE). To elaborate with, the fundamental thinking of a graph-based model is that of voting or recommendation. It is essentially casting a vote for the other vertex as one vertex connects to another one. The higher the number of votes cast for a vertex, the greater of the vertex's value. They can be utilized to identify a text or serve as a succinct description for a given document. Besides, for terminology extraction and the creation of domain-specific dictionaries, a method for the automated recognition of essential words in a text may be applied. Furthermore, for automatic summarization, the other *TextRank* application experimented consists of sentence extraction. In specific, the problem of extracting sentences can be considered like removing keywords, as both applications tend to find more 'representative' sequences for the given text. Consequently, two novel, unsupervised keyword and sentence extraction approaches were proposed, and evaluated and showed that the accuracy achieved by *TextRank* in these applications is comparable with that of state-of-the-art algorithms previously proposed.

In 2008, J. Herrera et al. emphasized the importance of the knowledge on statistical distribution of words throughout a text document. They used *spatial* statistical data analysis towards

detecting the most relevant words of a text by referring ranking system. Based on the Shannon's entropy of information, they proposed a new system for automatic keywords extraction from a text document [11]. To extend with, they have considered two indices like σ and Γ to find the how relevance among each word based on the previously applied application. However, they have improved the value of these measurements by observing the random distribution of the text over a text document. On the other hand, they have also introduced a new measurement unit named K_{nor} which helps to calculate the occurrences of the words based on the skewness of the distribution also introduced another index to extract keywords based on the information entropy. In this research, they applied *The Origin of Species* by Charles Darwin as a reference text data. However, the introduced method can be applied for anu natural language without requiring any previous knowledge like semantic or syntax similarity between the words.

M. Paukkeri et al. (2008), represented a new method named *Likey* [12] for key phrase extraction based on statistical data analysis where they claimed that it is a language independent method. In that research, the scholars have applied an external corpus as a reference dataset called *Europarl*, a natural language processing system, which seems a regular language format so that it is easier to adjust more than one language structures. Moreover, *Europarl* covers eleven European languages (French, German, Spanish, Italian, Portuguese, Danish, Swedish, Dutch and English) which means this *Likey* method would work for all of them languages. The experiment showed that Greek and Finnish languages are syntactically different that Romance and Germanic Languages. However, the proposed method gave a statistical result for all considered European languages. Moreover, the researchers claimed that this method needs a small number of steps for preprocessing and there is no require for any external knowledge like POS tagging.

P. Carpena et al., 2009, demonstrated a new technique which does not need any external knowledge or data source to classify the text [13]. In this research, the candidate keywords are gathered according to the top frequent words over the text document. Beside that they considered *spatial distribution* of words throughout the text and classify it so that most similar words gathered each other, and rest of the words are scattered in the text. To extend with, the keyword detection has been performed based on the cluster in the text where they emphasized that the operation cluster must be statistically significant though this is a common technique for all documents but for a small size of document like an article may not be good due to fluctuations of the words and they are in small frequency usually. By contrast, the statistically significant

depends on the frequency of the respective words. Therefore, the proposed method combined both statistical significance of words distribution as well as its frequency. Since there is no prior knowledge needed, this application is applicable for a single document only.

A new patent on rapid automatic keyword extraction [14] has been invented by S. J. Rose et al. in 2012. To illustrate, the candidate keywords are extracted by the inclusion of delimiters and stop words after parsing the document. Thereafter, candidate keywords have been ranked according to the calculation of their frequency over the document where couple of statistical functions were used like co-occurrence degree and cooccurrence frequency or both together. Therefore, top scored candidate keywords are being considered as the final keywords. On the other hand, the scholars have claimed that most of the keywords frequently contain multiple words but rarely contain standard punctuation, regular expression or stop words which are frequently appear over a document.

In 2015, S. Siddiqi et al. published a new method [15] which is unsupervised and domain independent. Moreover, they also claimed that this approach does not rely on any external corpus. Besides that, there are some statistical features have been applied in this approach like *term frequency and spatial distribution* of words in the document. The reason behind the choosing of term frequency, the researchers wanted to make sure that the extracted keywords are the frequency words over the document. On the other hand, the spatial distribution of words will help you to find the relevant words from each other which means the most relating words will be gather in a state and rest of the words will be scattered which are considered irrelevant words. As an experiment, they applied *Hindi* language which performed better by indicating good parameters as output of this research. However, this approach is applicable for light dataset, but it could be applicable for any languages.

S. Luthra et al. (2017), introduced a hybrid technique for extracting keyword [16] by splitting the text into multiple domains based on a master keyword template. In this research, they concentrated on the graphical view of words' frequency using *WordNet* which helps for better keyword selection. Beside that they also considered on another statistical term called TF-IDF to make sure that the selected keywords have high frequency over the document. To extend with, all words are checked how relevance of its own domain rather relevance in the whole document. Moreover, the graphical representation would provide the optimum keywords from the candidate

keywords list based on their co-occurrence among each other. Therefore, the graph will construct an algorithm by relating the similarity of words. So, the constructed algorithm can be representing like a combined method of splitting domain and graph algorithm for generating relevant keywords from a document. As future work, the researchers has indicated that using WordNet might help to find out the semantic words and how relating to each other so that the efficient keywords could be extracted from a document.

B. Armouty et al., 2019, focused on a specific language, Arabic document, to extract keyword [17] by concentrating statistical features especially the most common terms *tf-idf* and first occurrence have been considered in this research. On the other hand, after getting the candidate keywords, the *Support Vector Machine* has been applied to classify them. To extend with, in this process, the proper preprocessing is required because of the facilitating of clustering state where researchers have claimed that the output of the closeting depends on the quality of preprocessing. Before preprocessing, the text document has been parsed into words, numbers, and punctuation. However, there are seven steps in the preprocessing stage like tokenize the sentences, numbers are removed, punctuations are removed, stemming, splitting sentences, stop words removed, and number of tokens has been calculated. After preprocessing, the statistical features (tf-idf and first occurrence) have been implemented. To classify the words, Support Vector Machine classification has been applied for word distribution over the document. Besides that, the Radial Basis Function was used to kernelling the words' points. For experiment, scholars have gathered 844 documents which are constructed from Aljazeera's website. This process is known as supervised machine learning where their finding was 0.77 as precision, and 0.58 as recall.

2.2 Linguistic (POS Tagging and POS Patterns)

A. Gupta et al. (2014), proposed a new technique to extract keywords since word is the smallest part of a document [18] to represent the overall meaning immediately. Moreover, this is an unsupervised and domain independent approach where there is no prior knowledge needed to employ with any new document. This process has been done by using linguistic and statistical features. However, the overall methodology has been divided into two major modules like Keyword Extraction Module and Domain Extraction Module. In the first module, the web page is the input, and the extracted keywords are the output. Besides, in the second module, the

extracted keywords are sent to specify the correspondent domain. After this process, the extracted keywords and domain are stored into a repository. To extend with, the keyword extractions module is divided into four steps which are cleaning and initialization, stemming and stop word removing, candidate keyword determination using statistical features calculation (term frequency, title, meta-tag, URL, anchor text and highlighted words), and determining significant keywords using linguistic features calculation (POS value, first position and last position). The researchers claimed that after the experiment, the result outperforms the previously introduced methods. Besides that, the proposed method is useful for representing a web page by getting a small set of keywords.

A. Awajan, 2015, presented a new method based on Arabic language [19] to extract keywords. The author claimed that this study performed without any external dependency like corpus or training. The proposed system has been developed by combining the linguistic and statistical features. In the preprocess, the text is cleaned and filtered to get the actual information so that users can get into the roots and morphological patterns of the described words. After applying the cleaning process, the extracted words are clustered into equivalent classes to calculate the *derivate* and *non-derivate* words together. To identify the *n-gram* text, the author applied vector space model for defining the semantic similarity among the words so that most relevant words can be removed from the candidate keywords and unique informative words would be the expected words.

M. Rezaei et al., 2015, research's contribution is a new way of ranking the clusters that relies on the nouns' distribution over the text [7]. To begin with, the aim is to extract keywords with complete coverage of the page topics from the main text field in irrelevant text, such as short news articles on the news page. The proposed method is unsupervised, domain independent, not require corpus, and does not rely on HTML structure. Moreover, this research also focuses on studying the effects on the keyword extraction task of average-linkage, complete-linkage clustering, and the person assigned keywords. To extend with, the research methodology is divided into six segments, for instance, preprocessing, Parts of Speech (POS) tagging, lemmatization, similarity measure, clustering and selecting keywords. To extract text nodes from the tree, XPath1, which is a query language for addressing sections of an XML document, extracts symbols and numbers from the text, after which the length of - each node is computed. Authors extracted *unigram* nouns as candidate keywords by applying POS tagging to text fragments. In a phrase, sentence,

or paragraph, POS assigns sections of speech such as noun, verb, and adjective to each word in the text, based on its meaning, and relationship with adjacent and related words. Lemmatization is often helpful when the frequency of terms in a text is counted. Using the *Wu and Pulmer* test based on the WordNet, the researchers calculated the semantic similarity between all pairs of unique lemmas. On the other hand, hierarchical clustering has been considered because simple thresholding can regulate the number of clusters. If the similarity between their lemmas is greater than or equal to the threshold, the nouns are clustered together using an *agglomerative algorithm*. However, the frequency of nouns has been used as a criterion for selecting keywords from each cluster on the web page. Based on their frequency on the page, researchers rank the nouns in each cluster and pick the top frequent nouns. The best outcomes were addressed by clustering the nouns with the synonyms. Besides, the distribution of nouns around the page is more effective than the frequency of words.

T. Weerasooriya et al., 2017, demonstrated a new model for managing twitter data using Natural Language Processing (NLP) tools [20]. Due to the difficulties to handle tweets by NLP tools, they applied *Stanford CoreNLP Part-of-Speech (POS) tagger* to extract useful keywords from tweets. Moreover, this study was based on rule-based parsers and two external datasets have been used. However, this phenomenon has been done by two separated stages. In the first step, they processed the domain specific data after analyzing the tweets. In this state, the POS tagger extracts the keywords while parser helps to fetch the rest of the keywords if POS tagger misses something useful. After selecting the keywords, they have tested it by comparing with *CoreNLP POS tagger* using *Tuning test* strategy. The second step starts by executing *Named Entity Recognition and lemmatization*. In the second stage, the Tuning test has been implemented again on the extracted words to make sure that the keywords are meaningful. According to the researchers' data, the overall performance has been improved from 50% to 83.33%. However, this model gives better output only when it is applied for a domain specific data since the system has been developed based on NLP tools.

In 2019, H. Shah et al. introduced a new technique [21] of extracting keywords that apply the semantic similarity between the frequent terms on the web page and Parts of Speech (POS) tags' distribution. Moreover, hierarchical clustering is used by the writers of this study to cluster semantically related terms that have more coverage of the web page's content. Besides, there are two modules of this proposed research, namely preprocessing and extraction of keywords that

can be found. To extend with, the preprocessing module includes the extraction from the web page of a natural language document. The extraction module for keywords uses text from the preprocessing module. Then, the preprocessing module's first three functions include separating text from all the other content on a web page. All web page content is extracted using the Document Object Model (DOM) and X-path functionality. In the text filtering feature, text belonging to the JavaScript, scripting language and cascade style sheets are removed. Moreover, special characters such as @, *, £, or \$, punctuation marks, and numbers are also filtered out in the text filter function using a regular expression. On the other hand, the semantic similarity of two different words was computed in this study using path-similarity based on WordNet. The words that do not have WordNet synonyms are omitted from the list. In terms of their relation, the path-similarity metric measures the score between two separate words. On the other hand, for nouns, adjectives, and verbs, three matrices of similarity are generated separately. In clustering to the associated terms, similarity matrices are used. To locate the related terms in the lists, scholars use agglomerative clustering. After that, the clusters are scored by counting the frequencies in each cluster of all the terms. Based on the ratings, the clusters are ranked, respectively. By contrast, researchers used the noun list as the preliminary list of keywords for candidates. The most common words from the list of adjectives are added. The addition of just a single adjective gives the experiments highest accuracy and recall ratings. Similarly, following the inclusion of adjectives, top frequent verbs are added to the candidate keywords list.

On the other hand, Gagliardi et al. (2020) proposed a new method for unsupervised keyword extraction [22] where the authors have done some experiments on two specific languages, English, and Italian. In the process, there are two methods have been applied like word embedding models and clustering algorithm. For embedding words, they used *Word2Vec* and *GloVe* (pre-trained vector) which will provide the semantics similarities among the relevant words and their use of context. Moreover, the algorithm has been used for classifying the candidate keywords to detect the most significant words over the document which will help you to identify the real keywords. Though, the experiment shows a good output for English dataset but there are still some errors in the Italian language due to the use of Natural Language Processing features (language specific) and lack of grammatical structure. They have also mentioned the use of Wikipedia in the future to measure the similarity of words.

2.3 Structural (HTML and DOM Features)

S. Gupta et al., 2003, developed a framework [23] based on Document Object Model (DOM) tree rather concentrating only raw HTML elements. In analyzing section, at the very first, researchers carried out HTML parser to get the DOM tree presentation so they the targeted features could be detected easily since DOM tree represent the hierarchical view of the HTML document. After that, they filtered the contents by using various techniques to get the actual information. The plain text would be the output of this process. After analyzing the text content, keyword extraction process can be carried out. However, they claimed that their system gives better permeance by offering publicly accessible proxy to extract the content of a web page.

On the other hand, P. M. Joshi has published an article [24] for web text extraction in 2009. This research has been done based on DOM tree analysis as well as natural language processing. The researchers announced a general technique for web content extraction which does not depend on any external web template or corpus. They strongly believed that by extensive analyzing the web structure, the actual contents can be extracted. Beside that it is also necessary to employ the knowledge of natural languages process to be able to get the more accurate content. In the process, the web page or a list of web pages would be the input of this system and DOM tree representation will be the output. After that, research could easily get into the deep of the contents like block or sub-blocks and then body text, link or images could be identified easily. After this process, they applied semantic similarity algorithm to eliminate the relevant content which may be the unique content of the web document.

In 2011, L. Zhang et al. demonstrated a DOM-based algorithm [25] for extracting web information. The reason using DOM structure is that it would be easier to find the accurate information by searching nodewise since the current websites are standardized with DOM tree structure. Moreover, by detecting the XML documents they labeled the information to classify it. Besides, the researchers divided the overall process into three parts like characteristic selection and extraction, similarity calculation and extracting web page with multiple records. The whole process follows by the two strategies: comparison and classification. The experiment showed that this algorithm gives better performance where they applied it for two set of datasets, IMDB and OKRA, thought, the consistency of this result might very because of the dependency on DOM structure.

D-rank has been published in 2019 by H. Shah et al. regarding DOM based keyword extraction from web page [26]. In this research, a single web page is focused solely on the page's text and structural details. The Universal Resource Locator (URL) and other web page's positions, including the title, headings, and hyperlinks, provide useful information about the key terms on the web page, in addition to the term frequency. According to the words' location and their frequencies, various scores were assigned in this analysis. To elaborate with, the HTML content and URL for extracting candidate keywords were first preprocessed due to a web page including its address or Uniform Resource Locator (URL) and its content as an HTML file. Then, the scores are given in the different words based on the details from the HTML tags that define word positions. Finally, the process called D-rank, selects the top 10 candidates as representative web page keywords. The words are scored based on their positions on a web page: URL, title, six header levels, and hyperlinks that provide relevant keyword extraction information. Moreover, the term, frequency, has also been included in the scoring system. To identify candidate keywords, the method of this study involves the following tasks: extracting actual text from HTML content, cleaning text from symbols, tokenizing text with individual words, detecting text language and retrieving the list of stop words for the language, and eliminating stop words from the list of words. In the scoring process, the lists are used to add various scores to the words from various sections of the web page. To provide individual tokens or words, the researchers tokenized the resulting text fragments in the next stage. The process tokenizes the URL and at the same time, eliminates symbols and special characters, but removes the stop words from the list of words that are then carried out, which is an essential step in the method of keyword extraction.

By performing the literatures' study, we have been introduced with the various ideas and technical terms as well as the way to think in different ways how to research for a new problem. Besides that, it is clearly apparent that keywords extraction is one of the most important things in text processing, and still many things to do in this field. Many researchers have introduced numerous approaches how to handle the text document to extract keywords, key phrases, and sentences, but no one has concentrated on multilingual text document within a single web page which still appears in the web documents according to our survey. Therefore, we have introduced a new method to extract keywords from multilingual web pages by detecting languages.

CHAPTER 3. PROPOSED MULTILINGUAL KEYWORD EXTRACTION METHOD

3.1 New Approach for Extracting Keyword

There are numerous keywords extraction methods exists, but no one has considered regarding multilingual keywords extraction yet. In this research, we have focused on multiple languages keyword in a single web page which means there might have more than one language throughout the document. At first, we tried to analyze the DOM structure of the web page so that we can separate the text and other features of the structure. After that, the clean and filter technique has been applied to get plain text only so that we could detect the used languages which will help us to call relevant stop words. In our research, we considered two topmost used languages rather choosing all. The reason is that we may get plenty of languages during detection since sometimes similar words may come from different languages, which would not be fruitful for our research. At this stage, we have removed all stop words which cannot be considered as keywords because of the most repeated words for respective language. After that, the selective features have been scored based on predefined scoring formula. Therefore, the top ten (10) high scored words are considered as final keywords.

Moreover, the baseline of this approach is constructed from an exist DOM based method named *D-rank* which has been published by S. Shah et al. in 2019. Before sketching the current method, we got some motivation from that method since it has quite similar features rather considering one language at a time. To extend with, D-rank can handle one language at time which means multilingual web pages cannot be managed. Therefore, there might have some possibilities to be stop words in the keyword list. As we assume that, stop words cannot be keywords because of the most repetitive words over a document. To address this problem, we proposed the current method based on HTML structure like DOM tree representation as well as representative language detection. To implement the proposed method, we choose some DOM features from D-rank with some modification and introduce a new technique to handle the used languages in multilingual web pages.

3.2 DOM Structure

This is the starting phase of keyword extraction to parse HTML tags for separating text. DOM means Document Object Module which refers a web page as a tree presentation of HTML document so that we can easily trace the desired contents specifically. This is used particularly for analyzing web page by parsing the contents so that the separating of the text from html tags, CSS and scripts would be easier for next processing. As per our study, many researchers have applied DOM tree strategy for extracting web contents including keyword extraction [7] and title extraction [4, 29].

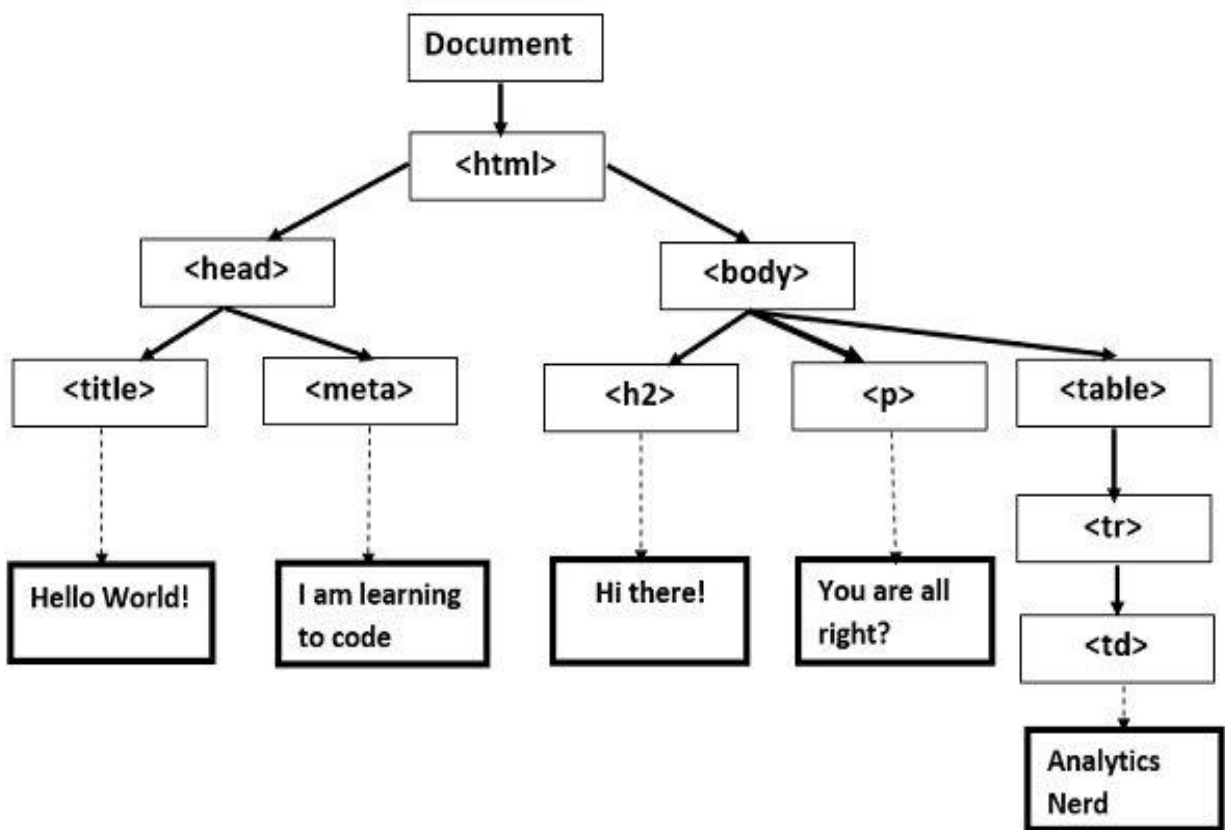


Figure 2. DOM Tree View

3.3 Feature Extraction

In this research, four DOM features have been chosen which are language independent to extract the valuable keywords as well as word frequency is considered which means most appeared words get higher score so that we could understand that these words would give important information towards getting good keywords. The features are title tag, headers (h1-h3), URL (host and path), and anchor (text label). After that, the scoring function has been done based on the different position of the page.

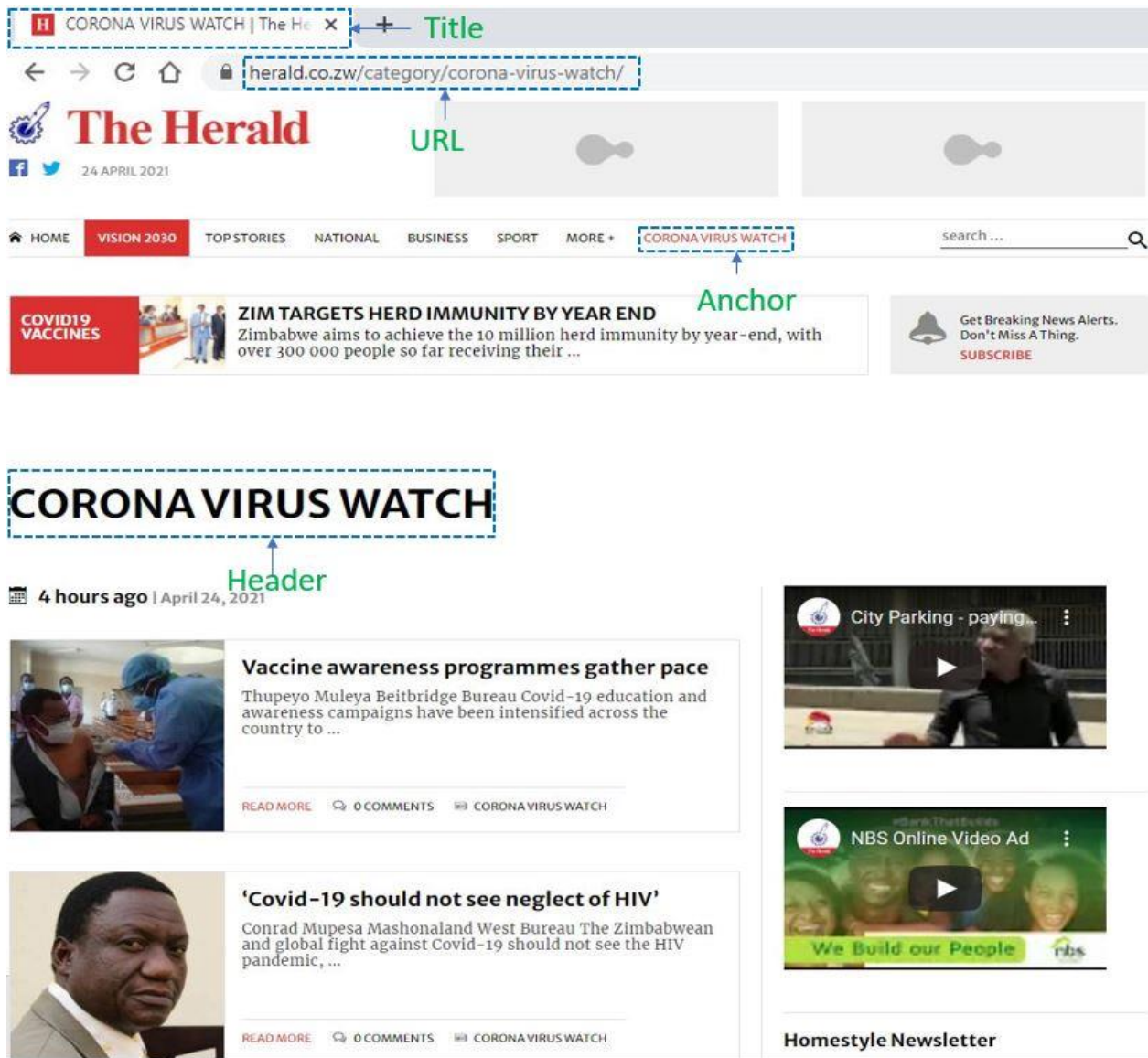


Figure 3. DOM Features

Title Tag

As we know that title tag is still the best source of getting important information which was the main motivation behind of this work. Particularly, title tag contains the most relevant information which could explain the whole document concisely. For example, in search engine optimization, usually the bot looks at the title tag to get inside of it for the valuable information which helps to manage in text processing to return relevant data to user. Therefore, we gave more importance on it by assigning higher score.

Headers (H1-H3)

Header is one of the most repeated uses throughout a web document of HTML tags which gives the text as extra highlighted view over the document. Usually, most important information is highlighted over a document to get users' attraction where headers play one of the essential roles. There are six (6) headers in HTML tags which are h1, h2, h3, h4, h5, and h6. However, according to the statistics, most of the extracted keywords come from the header one (h1), and then header two, three consecutively. Therefore, we have considered topmost three tags (h1 to h3) in this research where we give higher score for the header one and will decrease to header three.

URL (Host & Path)

URL refers Universal Resource Locator which addresses the website specifically over the world wide web as well resources. It has three sections like host, path, and query. Host refers the domain name like herald from <https://www.herald.co.zw> to specify and path refers the exact content what users look for. In this research we choose two of them (host and path). Moreover, most of the websites still bear essential information through the URL. For example, <https://www.herald.co.zw/category/corona-virus-watch/> which refers some vital information regarding the website. Therefore, URL could be a good option to get useful data for extracting keywords.

Anchor (Text Label)

Anchor tag also known as hyperlink in html document which helps to communicate among content to content or from one page to another page. Many researchers have applied this tag for content summarization or title extraction from web pages [27, 28]. In this research, we have considered only text level anchor to get useful information, for instance, `CORONA VIRUS WATCH`. From the following link, we can see that some important information is there which could be the user's keywords.

Term Frequency

The reason behind choosing this feature is to consider those words which are most repeated through the document so that user could assume that these are more important compared to other words. We could count how many times a word appears in the text document. Therefore, most repeated words have the priority to get highest score.

3.4 Text Extraction

As we know that a web page can be formed by the combination of different components like text, multimedia, html tags, styles, scripts and so forth. To get the raw text data from web pages, we can use DOM (Document Object Module) structure. By parsing DOM tree, the text and HTML features could be separated easily which will be ready for preprocessing.

Clean and Filter

In this stage, it is important to skip the unnecessary stuff from the document like HTML and CSS tags, punctuation, regular expression, and other scripts. This is also known as Tokenization where the targeted document is prepared for extraction by cleaning and filtering the unwanted components from the document. There are some open-source applications to analyze tags and scripts which produce clean text as an output. For punctuation and regular expression, we can

use some special functions to reduce them. Therefore, we could get it as plain text for the next operation.

Separate Text

The main reason behind splitting web text into sentences rather splitting by word is to detect the used languages more accurately since sometime same word might come from different languages. For instance, the word 'Radio' is used in various languages with the similar meaning like English, Finnish, Croatian, Dutch, European Spanish, Danish, French, German, Italian, Norwegian, Polish, Swedish, and Latin American Spanish [30].

3.5 Language Detection

Detecting the original languages is one of the challenging jobs in this research. As we have discussed earlier those similar words might come from different languages which may create conflict to researchers to take into consideration. Therefore, in this research, we have considered top two used languages (accuracy) to avoid complexity to handle all detected languages.

Moreover, it would not be wise decision to take into consideration for all detected languages since they might not be the original languages which are used in the respective web page. So, less used languages have less opportunity to be the accurate language therefore, it would be better to avoid to those language.

On the other hand, to detect any language there are plenty of methods available which are language specific. For example, Natural Language Processing (NLP) is one of the most popular ways to detect languages, but it can detect one language at a time from a text document. However, our goal is to detect multiple languages at a time to remove stop words from the text. For instance, a web text can be stated in mix mode languages so that we need to detect multiple languages at a time.

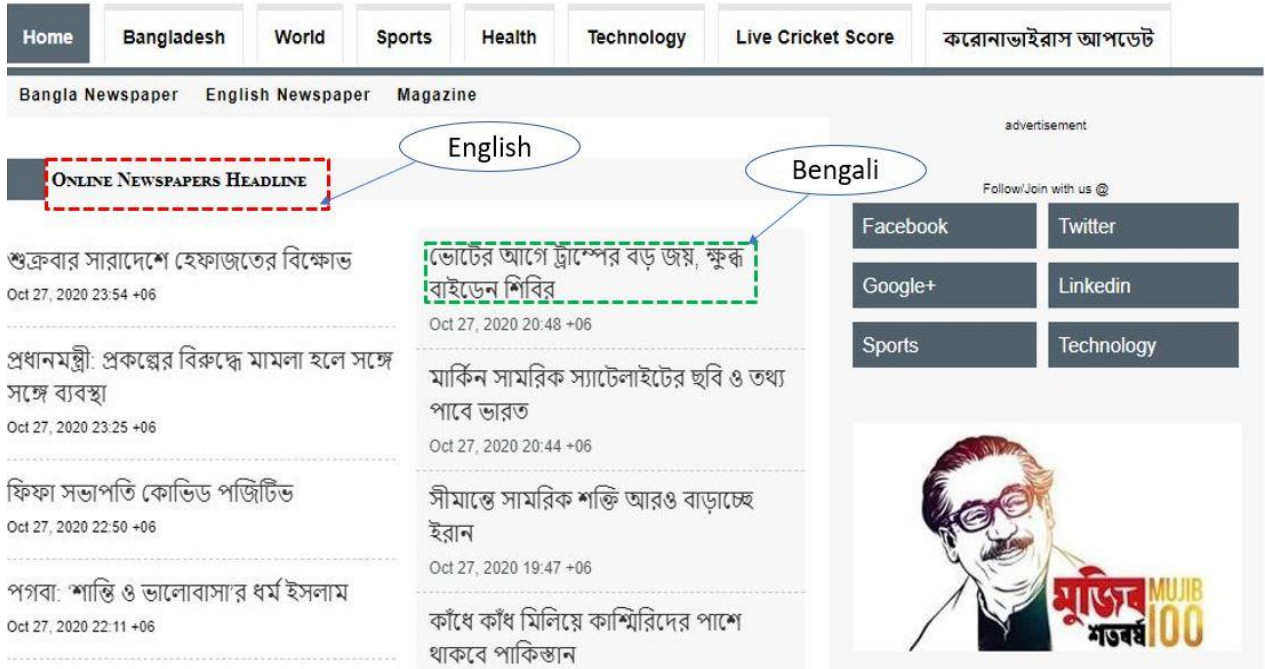


Figure 4. A Mixed-Mode Web Page

To extend with, in our experiment, we have chosen an open-source software package named *Polyglot* which is available for all users to use in online. The reason to choose this package is that it has wide range of language coverage with all possible languages including the percentage of accuracy so that we could choose topmost two used languages. In our observation, this package (*Polyglot*) supports around two hundred (200) languages which is the best supportive package as an open source. Besides that, the *Polyglot* provides some extra information including language code which is useful for us to detect multiple language at a time. This function is quite essential for the current system.

In Figure 5, an example of language detection has been demonstrated with different languages. We can observe that a list of possible languages is presented for a line of text with the confidence of all possible languages. However, we choose only two languages (topmost confidence) though there three languages appeared in the last example. So, most confidence value language has the most possible to be in the final language list.

খুলনা বিভাগে গত ২৪ ঘণ্টায় করোনায় ৪৫ জনের মৃত্যু হয়েছে

name: Bangla	code: bn	confidence: 99.0	read bytes: 528
name: un	code: un	confidence: 0.0	read bytes: 0
name: un	code: un	confidence: 0.0	read bytes: 0

Muktijuddho Mancha gives 48-hr ultimatum to arrest Whip Shamsul, Sharun

name: English	code: en	confidence: 98.0	read bytes: 617
name: un	code: un	confidence: 0.0	read bytes: 0
name: un	code: un	confidence: 0.0	read bytes: 0

Delhi Unlock 8: मॉल, सिनेमाघरों को मिली राहत, देखें क्या हैं स्कूल-कॉलेज के लिए निर्देश

name: Hindi	code: hi	confidence: 92.0	read bytes: 902
name: Khasi	code: kha	confidence: 7.0	read bytes: 1102
name: un	code: un	confidence: 0.0	read bytes: 0

Mumbai के Dharavi में Vaccine के लिए आधा किलोमीटर लंबी लाइन, Social Distancing का उड़ा मजाक

name: Hindi	code: hi	confidence: 71.0	read bytes: 794
name: English	code: en	confidence: 15.0	read bytes: 472
name: Sanskrit	code: sa	confidence: 5.0	read bytes: 1152

Figure 5. Language Detection

3.6 Stop Words Removal

Stop words means the most common words which are used frequently over a document. For example, a, an, the, he, has, had, have, am, are, was, were, and so on are the stop words which may not be the desired keyword. Therefore, stop words throughout the document needs to be removed to get the plain text for extracting keywords. This is because the stop word cannot be in the keywords list since it is the most repetitive words over a document. There are so many external resources such as libraries of stop word lists which could be used to remove them from the candidate keyword list.

On the other hand, there is a challenge how to get the stop words for the respective languages (more than one) at a time. Therefore, we have applied an open-source application for detecting the used languages and based on those languages the relative stop words could be called and remove from the text. To call the respective stop words, the detected languages have been combined and request for stop words together.

3.7 Keyword Selection

After removing the stop words, the candidate keywords are ready to be final keywords. In this state, the candidate keywords are scored with some numeric values to give the importance to be a keyword. Finally, top ranked candidate keywords are considered as the desired keywords. In our proposed method, we recommended to select the top ten best scored words are the final keywords though there is no obligatory to choose ten keywords only. It is just a standard to show off to other users for a general point of view.

3.8 Evaluation Measures

To evaluate the experiment, we defined some sort of metrics by giving numeric values which will give us visible result. This will help you out to find the most important keywords which could represent the respective web page to a new user. During setting score for different features, we just put some values based the importance throughout the page.

In Table 1, it represents the scoring process for proposed features where header one (H1) is scored by six (6), and rest of the headers (H2 & H3) valued respect wise by reducing one (1) in each step. Moreover, beside the headers we give importance on title tag which is scored by five (5). After that, the URL has been presented by putting value of five (5) and four (4) for host and path, respectively. At the very last, the anchor tag is given a little bit less importance by putting value of two (2).

Table 1. Scoring Metrics

Feature	Score
Header one (H1)	6
Header two (H2)	5
Header three (H3)	4
Title tag	5
URL (Host & Path)	5 & 4
Anchor	2

3.9 Workflow

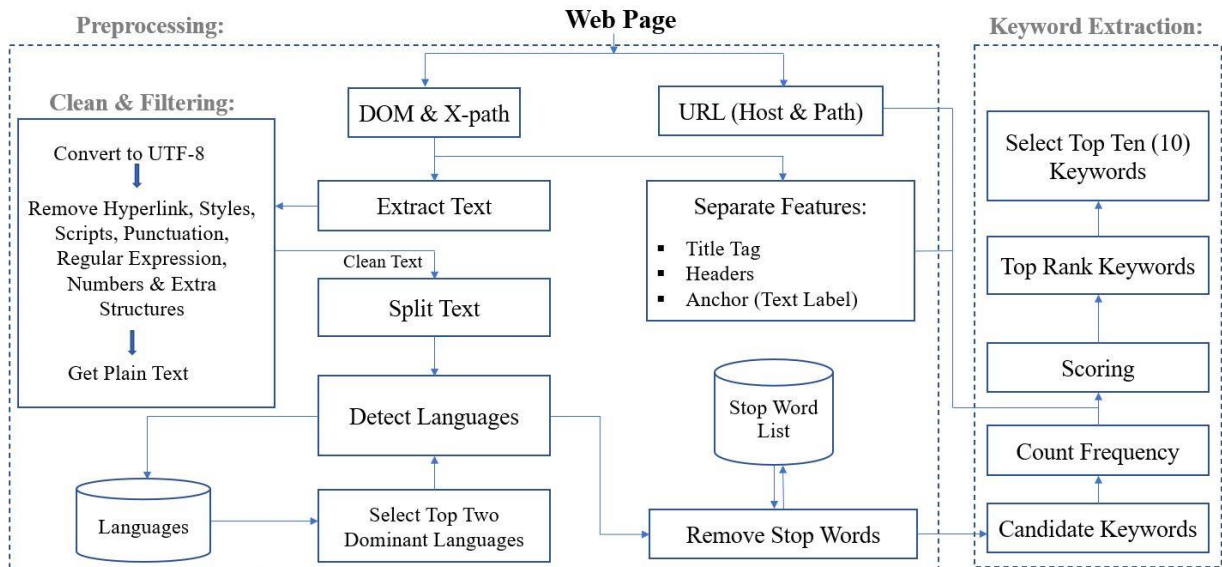


Figure 6. Workflow of The Proposed Method

CHAPTER 4. EXPERIMENT

Experimenting is one of the most common things for any research especially in computer science to defend the researcher's thought. Likewise, an experiment has been run to evaluate our proposed system. As an experiment, we have built a framework using some programming languages to test any sort of dataset regarding keyword extraction.

4.1 Data Collection

Data collection is one of the most important parts of this thesis. During data collection, we tried to make sure that every web page has at least more than one meta keyword as well as more than one language used so that we could realize that our proposed method would be good enough for all kind of web pages. Though, it was quite hard to find out such kind of web pages with the mentioned characteristics, fortunately we abled to manage at some point by spending bunch of time in web searching. To prepare the targeted corpus, we decided to collect hundred (100) web pages which could be acceptable in general for everyone. To elaborate with, to scrap the web pages we have used some open sources software to read the data especially meta keywords. If there is more than one meta data, the respective pages have been saved. Likewise, we choose hundred (100) multilingual pages.

In Figure 7, the meta keywords extraction process has been shown by indicating that meta keywords are from more than one language. Before extracting meta keywords, we searched randomly in Google for multilingual pages. Once we notice more than one language are appeared we immediately check whether it has more than one meta keyword or not using open-source meta tag analyzer (SEO Tool Center). For example, the website (abplive.com) has formed in two languages like English and Hindi, in where meta tags are also from double languages which gives us confirmation to store the respective web page into our directory.

Meta Tags Analyzer


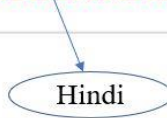
Page URL	Abplive.com/
Meta Title	Hindi News, Breaking News in Hindi, हिंदी न्यूज़ , Hindi Samachar, हिंदी समाचार, Latest News in Hindi, ताजा खबरें -ABP News <small>Ideally, your title tag should contain between 10 and 70 characters (Yours 123 characters)</small>
Meta Description	Latest News in Hindi, Hindi News Headlines, Breaking News in Hindi, हिंदी न्यूज़, ताजा खबरें, Hindi Samachar on ABP News. हिंदी समाचार, Latest News in Hindi from India and World on ABP News. <small>Meta descriptions contains between 160 and 320 characters (Yours 253 characters)</small>
Meta Keywords	ABP News, ABP न्यूज़, Breaking News, Breaking News in Hindi, ताज़ा हिंदी समाचार, हिंदी समाचार, Hindi News, News in Hindi, Latest News in Hindi
Meta Viewport	width=device-width, initial-scale=1 <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="text-align: center;">  <p>English</p> </div> <div style="text-align: center;">  <p>Hindi</p> </div> </div>
Open Graph	Open Graph meta tags is present

Figure 7. Meta Tag Analyzing

Moreover, we also downloaded the respective index page for future uses. The reason behind of this process is that web data might not be stable for all time which may affect the output of our experiment. Therefore, it would be wise decision to save the representative index pages.

To store the representative index page, we have used a third-party software named wget-1.20.3-win64 (a web scrapper) which will help you to download the desired page. Though, the success ration of this process is not up to the mark since most of the current websites (latest developed pages) are maintain security standard which does not allow for any external access. However, we have abled to find out of our targeted pages after spending bunch of time. Therefore, we stored hundred index pages into our storage for the next uses.


```
Wget https://www.desh.tv/education?start=180
Microsoft Windows [Version 10.0.19043.1083]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Omistaja>cd OneDrive
C:\Users\Omistaja\OneDrive>cd Desktop\wget-1.20.3-win64
C:\Users\Omistaja\OneDrive\Desktop\wget-1.20.3-win64>wget -k -K -E -r -l 10 -p -N -F --restrict-file-names=windows -nh https://www.desh.tv/
--2021-07-22 15:18:56-- https://www.desh.tv/
Resolving www.desh.tv (www.desh.tv)... 2606:4700:3030::6815:547, 2606:4700:3036::ac43:8527, 104.21.5.71, ...
Connecting to www.desh.tv (www.desh.tv)|2606:4700:3030::6815:547|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: 'index.html'

index.html           [ <=> ] 324.41K  1.41MB/s  in 0.2s

Last-modified header missing -- time-stamps turned off.
2021-07-22 15:18:57 (1.41 MB/s) - 'index.html' saved [332200]

Loading robots.txt; please ignore errors.
--2021-07-22 15:18:58-- https://www.desh.tv/robots.txt
Reusing existing connection to [www.desh.tv]:443.
HTTP request sent, awaiting response... 200 OK
Length: 923 [text/plain]
Saving to: 'robots.txt'

robots.txt          100%[=====] 923  --.-KB/s  in 0s
```

Figure 8. Index Page Scrapping

After selecting the web pages, we parsed it using a web scrapper (BeautifulSoup – one of the renown python-based packages). For future uses, we split the pages in different portions like respective URL, HTML document, HTML features, plain text, meta keywords, and so on. Therefore, it would be easier to handle to web pages in different purposes.

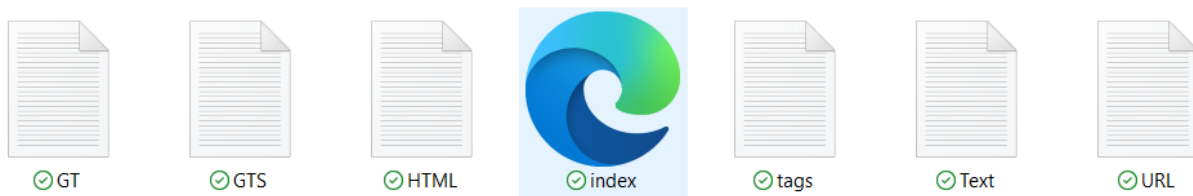


Figure 9. Web Page Splitting

Table 2. Dataset Properties

Dataset	Language	Size	Keyword	Location
Multilingual	More than one language in each web page	100	2-15	Web search

In Table 2, the properties of the created dataset have been demonstrated. To extend with, the dataset category was multilingual which means each web page must be formed in more than two languages. The size of the dataset was hundred (100). This is just a number to maintain the standard as well as for good analytical result. On the other hand, the number of meta keywords was more than two as well as more languages have been appeared in the meta keyword list. However, the whole phenomena have been done in online environment particularly randomly web searching.

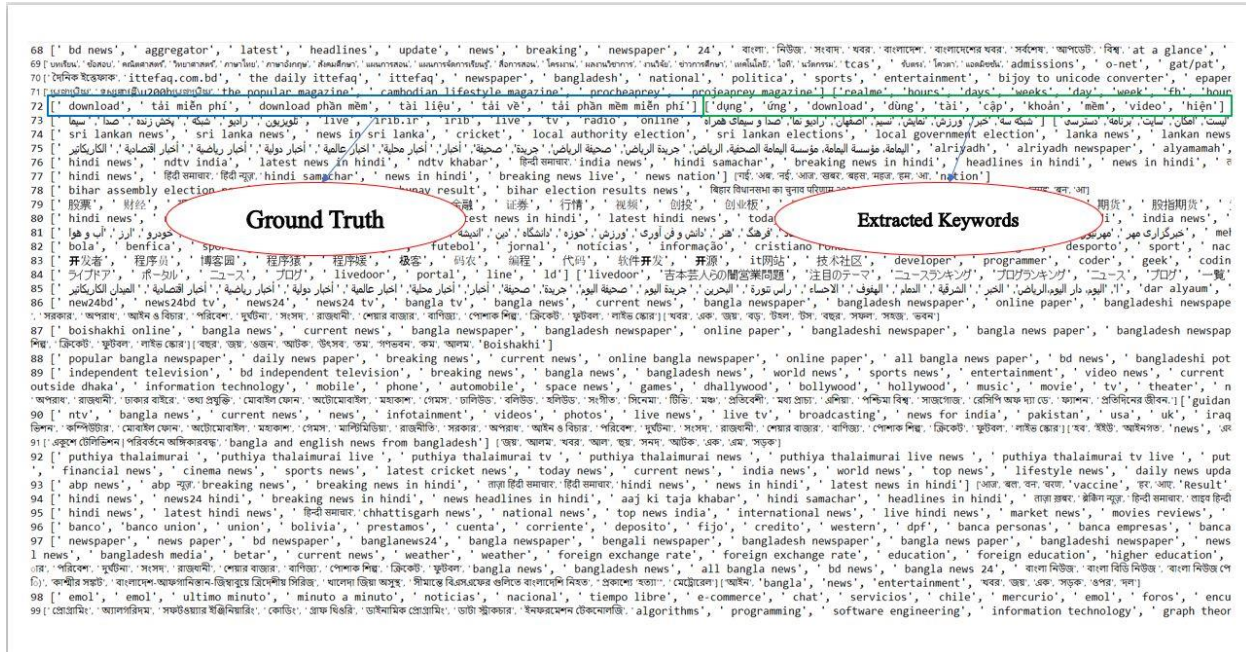


Figure 10. A View of The Dataset (Meta Keywords and Extracted Keywords)

In Figure 10, a view of the extracted keywords has been represented where meta keywords and extracted keywords from the proposed method are placed respectively. In details, there are three columns in the dataset in where the first column represents the index number, the second column refers the meta keywords portion which are placed from representative web pages and the third column presents the extracted keywords from S-rank method. Moreover, the meta keyword and extracted keywords are separated by a third parenthesis which will help to measure the accuracy of the performance of current system. Therefore, we have prepared three similar files for TextRank, S-rank and D-rank to compare the output for each method based on the multilingual dataset.

4.2 An Overview of Keyword Extraction Process

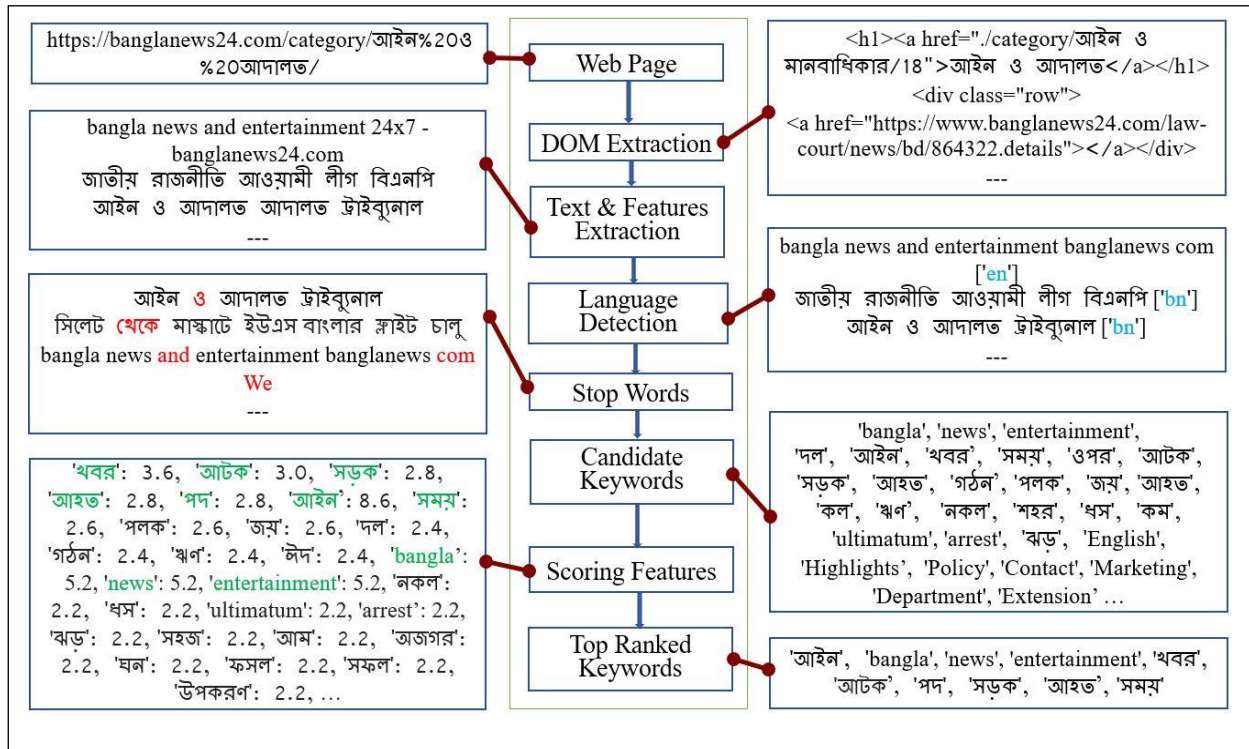


Figure 11. Example Workflow

4.3 Keyword Extraction Steps

To evaluate our proposed system, we have done some sort of experiments by engaging multiple lingual web pages. Here, we demonstrated an example with indication of some elementary steps which are described below:

Step 1

We choose a random domain (<https://banglanews24.com/category/আইন%20ও%20আদালত/>) where it is divided into two sub-sections such as host (<https://banglanews24.com/>), and path (<category/আইন%20ও%20আদালত/>).

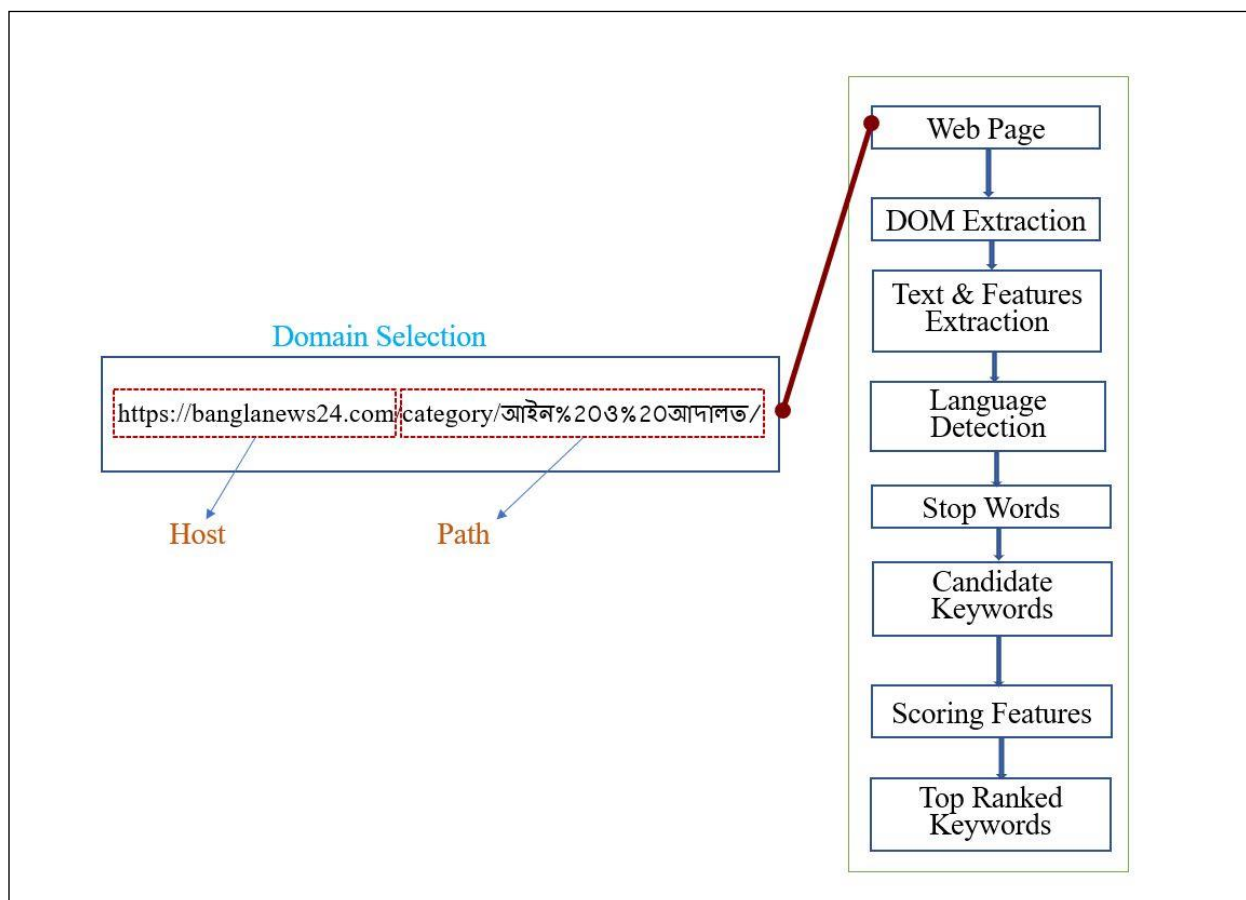


Figure 12. Domain Selection

Step 2

To extract the real information, it is necessary to investigate the whole HTML features (root of HTML). DOM tree representation is the actual way to reach all HTML structures. Detecting the DOM tree presentation, we applied a third-party API named BeautifulSoup “a python-based web parser” which allows us to read the HTML tags and other features.

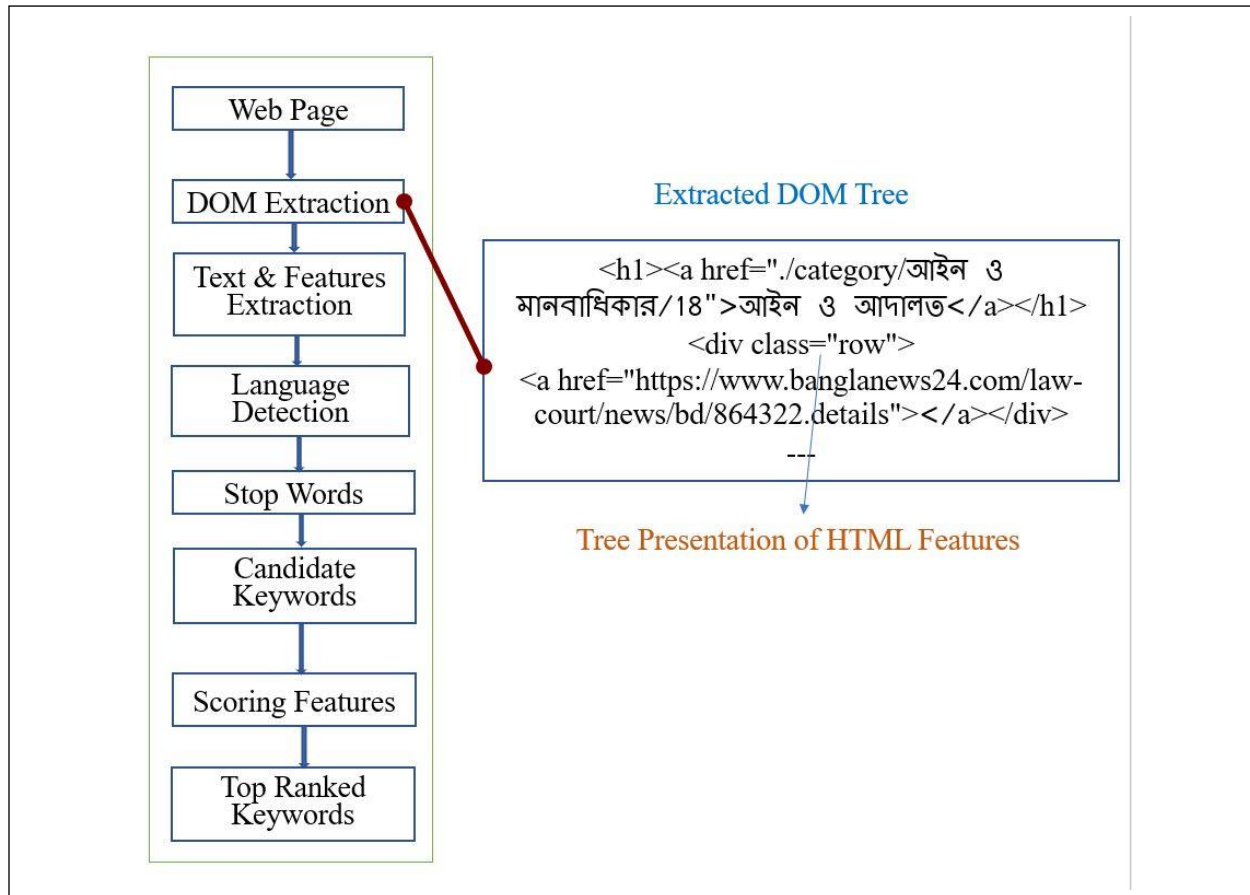


Figure 13. DOM Tree Presentation

Step 3

By the blessing of web parser like BeautifulSoup we abled to extract the text from the body and other sources of information in HTML. After extracting the text, we filtered and cleaned the text to get the plain text for next processing. This step is called preprocessing where the unnecessary elements are removed from the text like regular expression, numeric, scripts, tags and so on.

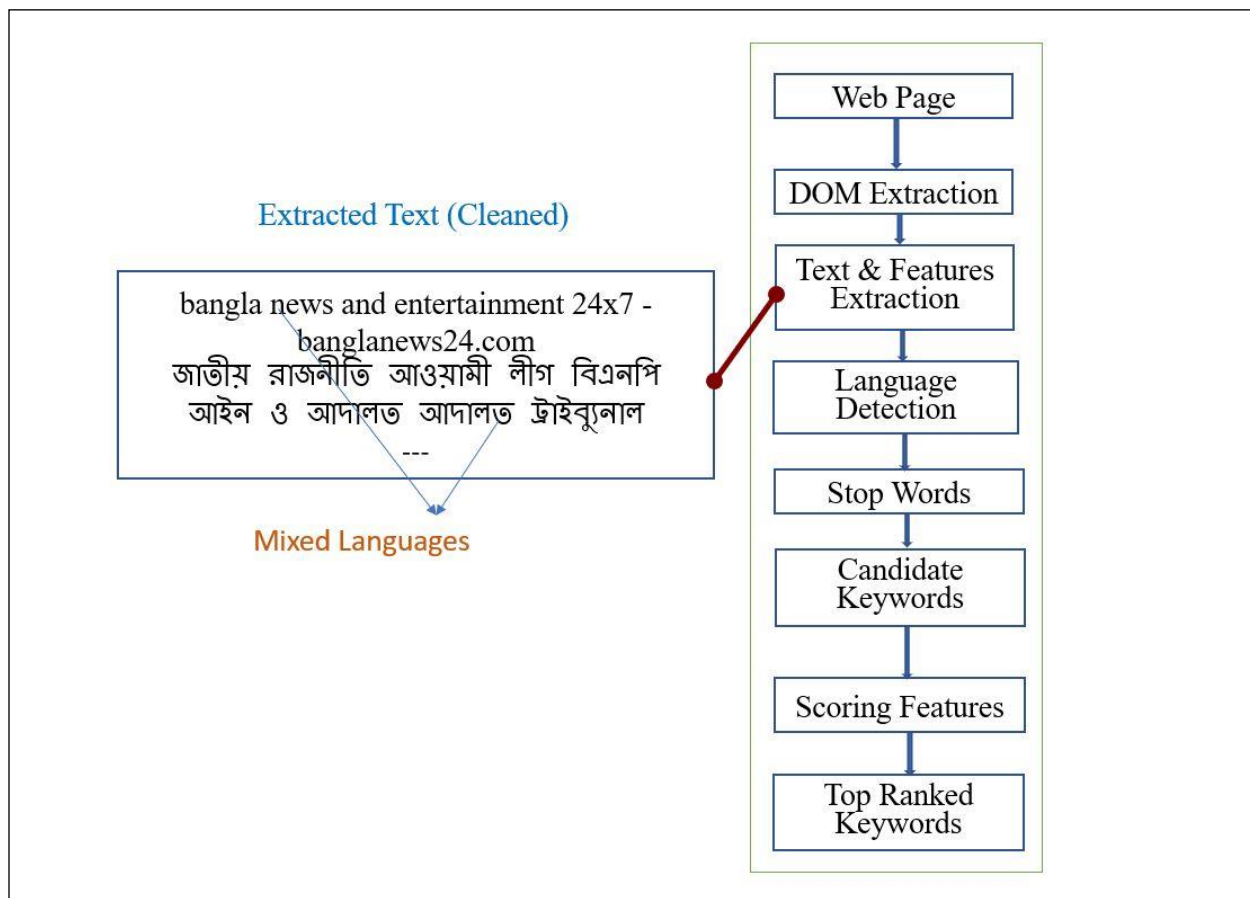


Figure 14. Cleaned Text

Step 4

In this step, we look forward for the stop words list from the repository so that we could easily detect the matching words in the text and remove it immediately. In order to get the double languages stop words list at a time, we used python query by combining the detected lanaguages (English and Bengali). In Figure 11, the red marked words are stop words which are revomed from the text to get the actual informative keywords.

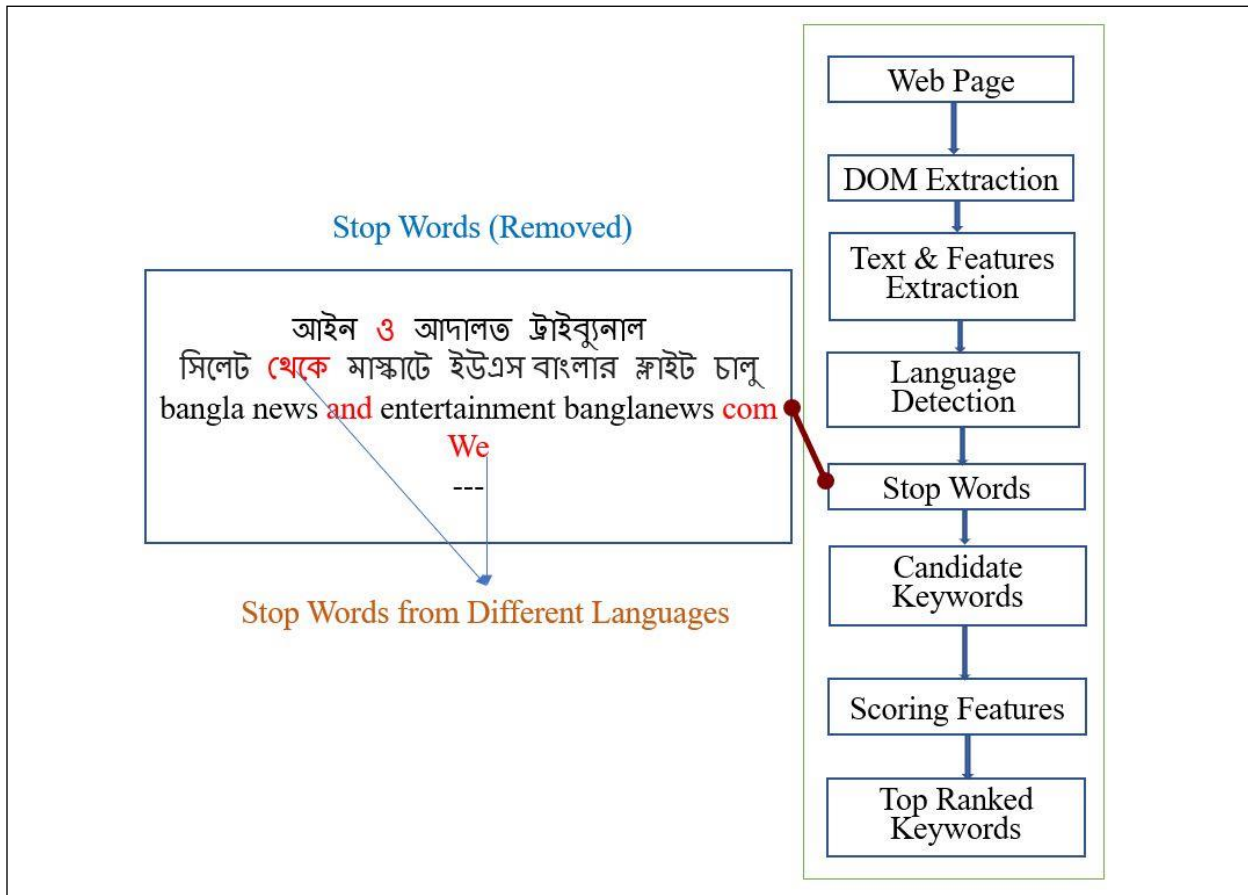


Figure 15. Stop Words Deletion

Step 5

After removing the stop words, the list of words which are considered as candidate keywords which could claim for targeted keywords. In Figure 12, we could see that there is a list of cleaned words which are ready to be scored and selection as a keyword.

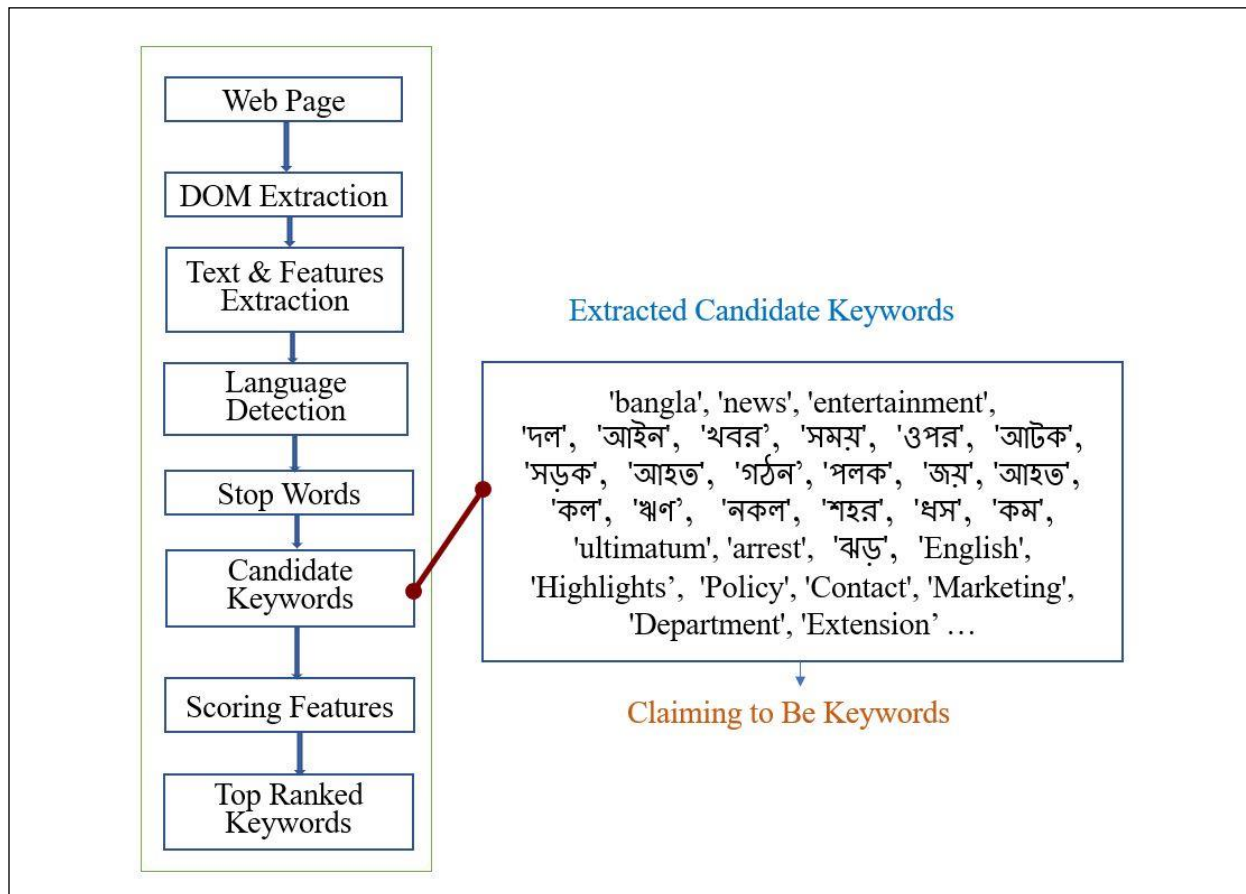


Figure 16. Candidate Keywords

Step 6

Before finalizing the keywords, we scored the candidate keywords to find out the best described keywords which could represent the whole text to a new user. According to the scoring criteria which is mentioned in the method description section, the top scored words are considered as keywords.

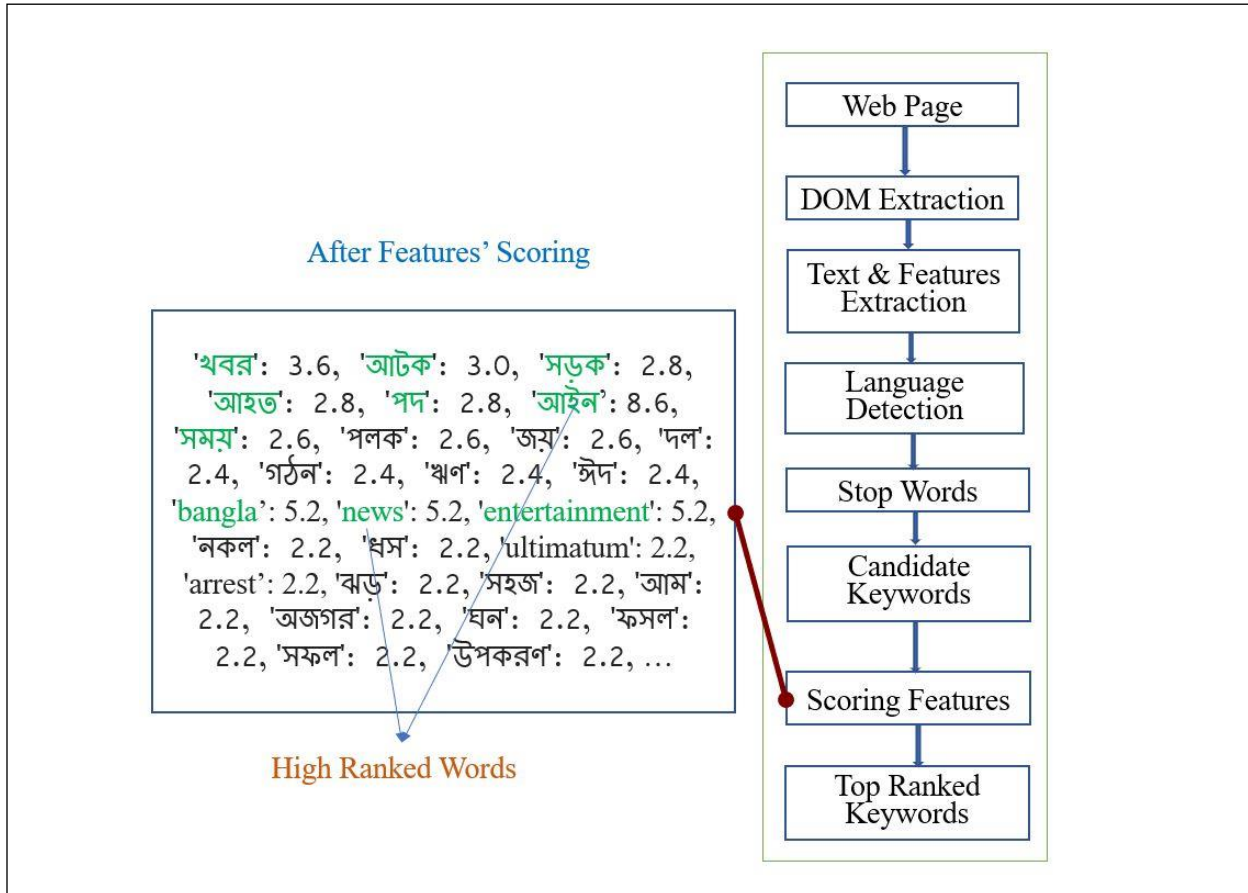


Figure 17. Scoring Candidate Keywords

Step 7

This step is called keyword selection where top ten scored are selected for the final keywords. The reason behind showing ten keywords is just a standard number to make showing it off for general view of users. However, we could notice that the output is also in mix-mode lingual keywords which gives us confirmation that our proposed system works for multilingual web pages.

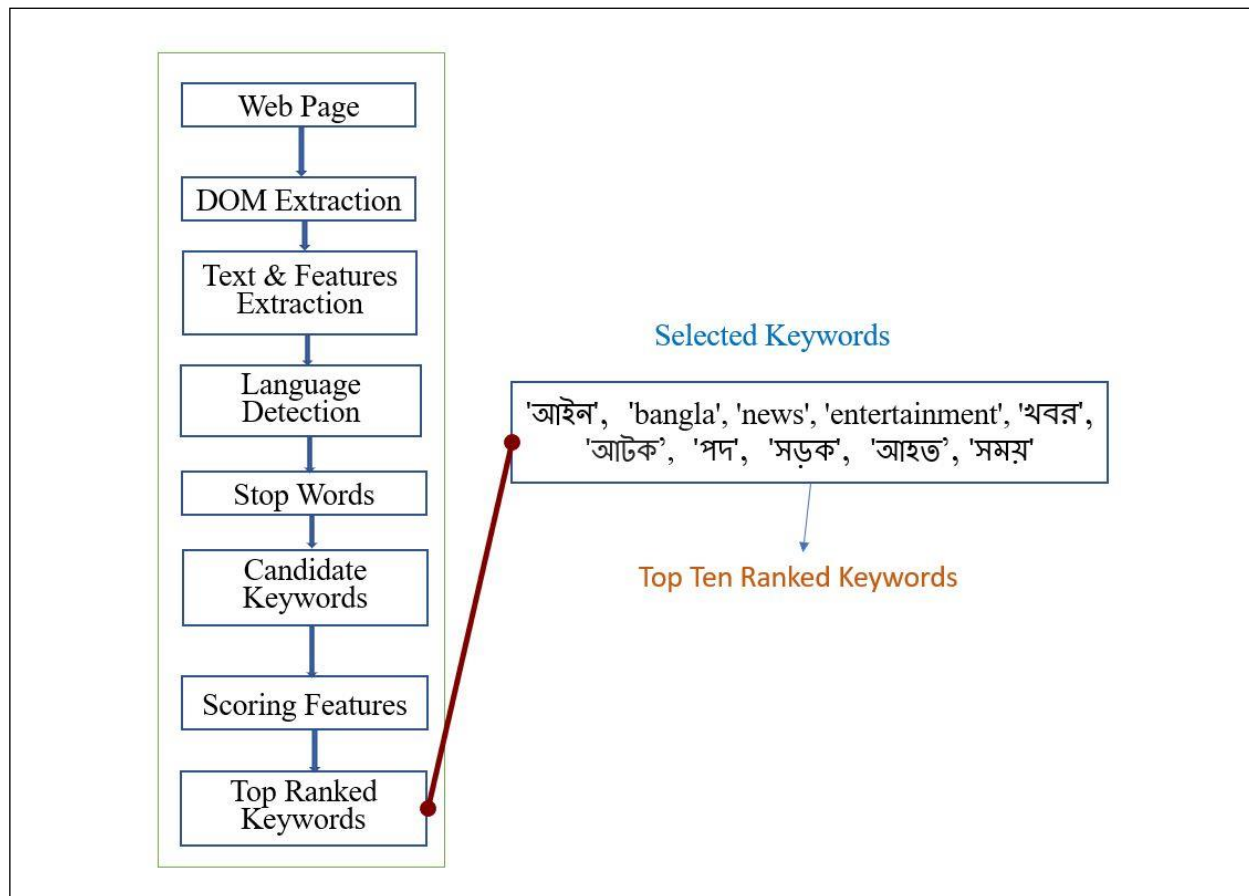


Figure 18. Keyword Selection

4.4 Technical Tools

During experiment we used some different programming languages, framework, API, and techniques, and so on which are listed below as:

- We developed an own framework
- Language: Python
- API: BeautifulSoup DOM parser, Polyglot, python packages & wheelers
- Software: wget-1.20.3-win64
- Meta tag analyzer: SEO Tool Center
- IDE: Anaconda, & Spyder
- Testing: test for multilingual webpages

CHAPTER 5. RESULT DISCUSSION

In Table 3, the presented data indicates the difference between D-rank and S-rank where D-rank is a DOM based keyword extraction method introduced by Himat et al. in 2019 and S-rank refers the current proposed method for multilingual web pages. Moreover, D-rank can perform for one language at a time which means if the representative website is formed by multiple languages, there are some possibilities to exist some stop words in the keywords. This is because D-rank performs only for the topmost dominant single language. As an experiment, we have run both program for a specific web page (banglanews24.com) which is developed in two languages like English and Bengali. In the final keyword list, we can clearly notice that D-rank still produces two stop words which are from Bengali language whereas S-rank provides quality keywords. By contrast, the TextRank can only perform for English languages which means it ignores the second language. However, the proposed method (S-rank) can perform for multilingual web pages which has been showed in the experiment. In Table 3, the last most column represents the S-rank's output where mixed-mode words are demonstrated in the final keyword list.

Table 3. A Comparative Output between D-rank and S-rank

Webpage	Ground Truth	D-rank	S-rank
https://www.banglanews24.com/	['bangla', 'news', 'bangladesh business', 'করোনাভাইরাস', 'রোহিঙ্গা', 'ফুটবল', 'bangla news', 'bangla news 24', 'breaking news of bangladesh', 'today's news', 'awami league', 'অসুস্থ', 'নিহত', 'মেট্রোরেল']	['ঈদ', 'english', 'আইন', 'highlights', 'bangla', 'news', 'entertainment', 'খবর', 'এই', 'সব']	['ঈদ', 'আইন', 'bangla', 'news', 'entertainment', 'খবর', 'জয়', 'সফল', 'আহত', 'সড়ক']

To evaluate the proposed system, we have done some experiments by engaging real web pages. To measure the success of an experiment, there are some statistical based functions which could give you a visible view regarding the experiment. In our test, we choose the statistical functions named precision, recall and f-score to see the accuracy of the experiment so that user could have some understanding how effective our proposed method is. To explain with, the term precision represents the percentage of the selected keywords which are real keywords. It is also known as positive predictive value. On the other hand, the term recall as sensitivity in diagnostic binary classification which demonstrates the percentage of the keywords which are selected. The final term f-score is the harmonic mean of the precision and recall which means the average weighted values of the precision and recall.

Moreover, the value of these terms could be zero (0) to one (1) where one (1) represents the highest performance of the test. For example, the highest possible f-score value one (1) indicates the perfection of precision and recall, and the lowest value zero (0) can only be assigned once precision and recall is zero. However, these measurements criteria are calculated based on the true positive, false negative, and false positive which are define as follows:

- True Positive = Number of correctly predicted values (keywords)
- False Negative = Number of missing values (keywords)
- False Positive = Number of incorrectly predicted values (keywords)

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

Table 4. Evaluation Measurements Values

Method	Precision	Recall	F-score
TextRank	0.21	0.06	0.08
D-rank	0.22	0.06	0.09
S-rank	0.26	0.08	0.10

During experiment period, we have evaluated for three methods using multilingual dataset. The methods are TextRank (one of the renowned methods in the keyword extraction field), D-rank (baseline method) and S-rank (proposed method). However, the experiments clearly indicate that the proposed method works better than other methods for mixed-mode websites.

In the experiment table (Table 4) shows the representative data based on the performance of the test. To extend with, the heading precision refers the values of 0.26 which means the percentage of detected keywords is 26% where the missing keywords is 8% in compared to the ground truth meta keywords for the proposed method whereas 0.21 and 0.22 are the precision of TextRank and D-rank respectively, and 0.06 as recall for both. Therefore, the success ratio of the proposed method 10% whereas 8% and 9% are TextRank and D-rank respectively. However, the hard evaluation where the actual ratio could be little bit higher. This is because, in hard evaluation, only the exact matching extracted words are to be considered with the ground truth which does not reflects the semantic similarity among words. For example, the word ‘entertainments’ would get zero (0) as precision value though it is mostly relevant to the extracted word ‘entertainment’.

However, in future, we could run a soft evaluation test using Levenshtein distance to measure the similarity between words which could give us better result since in the soft measurement, the correlation between words is measured so that the ration of actual keywords will be higher compared to the hard evaluation. Moreover, adding WordNet or Wikipedia could give us more semantic similar words to find the quality keywords.

CHAPTER 6. CONCLUSION

The amount of online data is being increased rapidly due to the climbing of virtual engagement in all sectors. It is also necessary to manage the heavy amount of online data for optimum utilization. Though, a bunch of research have been done on it but still it has some scopes for more investigation, for instance, there is no method which could handle to multilingual web document. To address this problem, we proposed a new method based on structural features of HTML (DOM tree representation). In the process, we parsed the HTML structure to separate the text and other features. After the preprocessing, the applied languages have been detected by line wise, which helps to find the related stop words and remove them from the extracted text. Finally, top ranked ten words are considered as keywords. As an experiment, we have prepared an own made dataset by collecting hundred web pages where each page has mix mode text. However, the proposed method works for multilingual web page at a time but still there are some limitations since it depends on the DOM structure where information might be changed at any time which may affect the accuracy of the output. Therefore, it is recommended to save the source data during the dataset creation.

Moreover, the result of the experiment shows that the proposed method works better for multilingual dataset. In hard evaluation, the overall performance (f-score) is 10% which is better than the renowned TextRank and base method D-rank. Therefore, we can claim that the current solution is fruitful for mixed-mode web pages.

Nowadays, keyword extraction has plenty of applications in different areas of information technology such as web data mining, improving the search engine performance, social network, location-based application, topic detection, indexing, classification, summarizing, content-targeting advertising and so on. For example, proper categorization of articles could help the search engine to retrieve optimal data to users as well as users could get an overview of any article by viewing the respective keywords.

REFERENCES

- [1] M. Grineva, M. Grinev & D. Lizorkin, 2009, Extracting key terms from noisy and multi-theme documents, *Proceedings of 18th International Conference on World Wide Web*, ACM New York, NY, USA, pages 661–670.
- [2] J. Feather & P. Sturges, 2003, *International Encyclopedia of Information and Library Science: Routledge*.
- [3] R. Mihalcea & P. Tarau, 2004, TextRank: Bringing order into text, *Proceedings of Conference on Empirical Methods in Natural Language Processing*.
- [4] N. Gali & P. Fränti, 2016, Content-based title extraction from web page, *Proceedings of International Conference on Web Information Systems and Technologies (WEBIST)*, volume 2, pages 204-210.
- [5] P. D. Turney, 2003, Coherent keyphrase extraction via web mining, *ArXiv Preprint*, cs/0308033.
- [6] P. Tonella, F. Ricca, E. Pianta & C. Girardi, 2003, Using keyword extraction for web site clustering, *Proceedings of 5th IEEE International Workshop on Web Site Evolution*, pages. 41-48.
- [7] M. Rezaei, N. Gali & P. Fränti, 2015, CL-Rank: A method for keyword extraction from web pages using clustering and distribution of nouns, *IEEE/WIC/ACM International Conference on Web Intelligence, and Intelligent Agent Technology (WI-IAT)*, pp. 79-84.
- [8] M. Chen, J. T. Sun, H. J. Zeng & K Y Lam, 2005, A practical system for keyphrase extraction for web pages, *Proceedings of CIKM*.

- [9] W. Yih, J. Goodman & V. R. Carvalho, 2006, Finding advertising keywords on web pages, *WWW 15th International Conference on World Wide Web*, New York, NY, ACM, pages 213–222.
- [10] K. S. Kuppusamy & G. Aghila, 2012, A model for personalized keyword extraction from web pages using segmentation, *International Journal of Computer Applications*, Volume 42, No.4.
- [11] J. P. Herrera & P. A. Pury, 2008, Statistical Keyword Detection in Literary Corpora, *EDP Sciences*, Societ`a Italiana di Fisica, Springer-Verlag.
- [12] M. Paukkeri, I. T. Nieminen, M. Pollä & T. Honkela, 2008, *COLING, 22nd International Conference on Computational Linguistics*, Posters Proceedings, Manchester, UK, pp. 18-22.
- [13] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado & J. L. Oliver, 2009, Level statistics of words: Finding keywords in literary texts and symbolic sequences, *The American Physical Society*, Volume 79, Issue 3, pp 035102-4.
- [14] S. J. Rose, W. E. Cowley, V. L. Crow & N. O. Cramer, 2012, RAPIDAUTOMATIC KEYWORD EXTRACTION FOR INFORMATION RETRIEVAL AND ANALYSIS, *United States Patent*, Patent No. US 8,131,735 B2.
- [15] S. Siddiqi & A. Sharan, 2015, Keyword and keyphrase extraction from single Hindi document using statistical approach, *2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India.
- [16] S. Luthra, D. Arora, K. Mittal & A. Chhabra, 2017, A Statistical Approach of Keyword Extraction for Efficient Retrieval, *International Journal of Computer Applications*, Volume 168, No.7.

- [17] B. Armouty & S. Tedmori, 2019, Automated Keyword Extraction using Support Vector Machine from Arabic News Documents, *IEEE Jordan International Joint Conference on Electrical Engineering, and Information Technology (JEEIT)*, Amman, Jordan.
- [18] A. Gupta, A. Dixit & A. K. Sharma, 2014, A Novel Statistical and Linguistic Feature Based Technique for Keyword Extraction, *International Conference on Information Systems and Computer Networks (ISCON)*, Mathura, India.
- [19] A. Awajan, 2015, Keyword Extraction from Arabic Documents using Term Equivalence Classes, *ACM Transactions on Asian and Low-Resource Language Information Processing*, Volume 14, Issue 2, No. 7, pp 1–18.
- [20] T. Weerasooriya, N. Perera & S.R. Liyanage, 2017, An Essential Keywords Extraction Model for Twitter Designed using NLP Tools, *Proceedings of the 10th KDU International Research Conference*, Sri Lanka.
- [21] H. Shah, M. U. Khan & P. Fränti, 2019, H-rank: a keywords extraction method from web pages using POS tags, *IEEE International Conference on Industrial Informatics (INDIN)*, Helsinki, Finland.
- [22] I. Gagliardi & M. T. Artese, 2020, Semantic Unsupervised Automatic Keyphrases Extraction by Integrating Word Embedding with Clustering Methods, *Multimodal Technologies and Interaction*, Volume 4, No. 2, <https://doi.org/10.3390/mti4020030>.
- [23] S. Gupta, G. Kaiser, D. Neistadt & P. Grimm, 2003, DOM-based Content Extraction of HTML Documents, *Proceedings of the 12th international conference on World Wide Web*, pp 207-214.
- [24] P. M. Joshi & S. LiuWeb, 2009, Document Text and Images Extraction using DOM Analysis and Natural Language Processing, *The 9th ACM Symposium on Document Engineering*, Munich, Germany.

- [25] L. Zhang, M. Li, N. Dong & Y. Wang, 2011, An Improved DOM-based Algorithm for Web Information Extraction, *Journal of Information & Computational Science*, Volume 8, No. 7, pp 1113–1121.
- [26] H. Shah, M. Rezaei & P. Fränti, 2019, DOM-based Keyword Extraction from Web Pages, *International Conference on Artificial Intelligence Information Processing and Cloud Computing (AIIPCC)*, pp. 1-6.
- [27] D. B. Bracewell, F. Ren & S. Kuriowa, 2005, Multilingual single document keyword extraction for information retrieval, *Proceedings of NLP-KE*.
- [28] G. Matösević, 2015, Using anchor text to improve web page title in process of search engine optimization, *Proceedings of Conference on Information, and Intelligent Systems*, Varždin, Croatia.
- [29] N. Gali, R. Mariescu-Istodor & P. Fränti, 2017, Using linguistic features to automatically extract web page title, *Expert Systems with Applications*, vol. 79, pp. 296-312.
- [30] Collins, 2021, Similar words used in different languages, *Retrieved from Collins Dictionary*, <https://www.collinsdictionary.com/dictionary/french-english/radio>.