Journal of Namibian Studies, 34 S2(2023): 743–755 ISSN: 2197-5523 (online)

Keyword/ Keyphrase Extraction from Text of Indian Election Domain

Seema Shukla [1], Apekshita Aggarwal [2], Tanishka [3], V N Shukla [4]

[1] Dronacharya Group of Institutions, Greater Noida, India

seema.shukla@gnindia.dronacharya.info

[2] Dronacharya Group of Institutions, Greater Noida, India

apekshita.16040@gnindia.dronacharya.info

[3] Dronacharya Group of Institutions, Greater Noida, India

tanishka.16183@gnindia.dronacharya.info

[4] Director IT (Retd.), Election Commission of India, New Delhi , India

vnshukla@gmail.com

Abstract

In recent years, there has been a tremendous increase in the amount of data generated from various sources including social media, news articles, and blogs. With the rise of social media platforms, people are expressing their opinions more freely than ever before. This has led to an explosion of data in the Indian Election domain, where people express their views on various political parties and candidates. In order to extract meaningful information from this vast amount of data, it is important to identify and extract relevant keywords and phrases. Keyword and phrase extraction is the process of automatically identifying important words and phrases from a piece of text. This process is crucial for various natural language processing tasks such as text mining, sentiment analysis, topic modeling, and text classification. In this research paper, we focus on the task of keyword and phrase extraction from Indian Election domain text. We aim to extract relevant keywords and phrases that are most commonly used in the context of Indian elections. This research is important as it can help in understanding the key issues and concerns of Indian voters during the election season. We use various natural language processing techniques and machine learning algorithms to extract keywords and phrases from a large corpus of Indian Election domain text. Our approach involves pre-processing the text, including tokenization, stop-word removal, stemming, and POS tagging. We then use various statistical and machine learning models to identify the most relevant keywords and phrases.

Keywords: Keyword and phrase extraction, natural language processing, text mining, unsupervised keyword extraction

1. INTRODUCTION

The Indian elections are one of the largest democratic exercises in the world, with over 900 million eligible voters. Social media platforms such as Twitter, Facebook, and WhatsApp have become increasingly popular among Indian citizens, politicians, and journalists to share their opinions, viewpoints, and news updates during elections. This has resulted in a massive influx of textual data, making it difficult for researchers to identify relevant keywords and keyphrases. Keywords and keyphrases are crucial in information retrieval, natural language processing, and data mining. Identifying relevant keywords and keyphrases from textual data can help in summarizing the content, understanding the sentiment, and classifying the data. In this paper, we propose an approach to identify keywords and keyphrases from Indian election domain text using machine learning techniques. In order to incorporate domain specific knowledge in the extraction process organizational and law named entities were identified and the candidate keys with higher similarity to these named entities were assigned a higher score thereby increasing the probability of their inclusion in the finally generated keywords and phrases. The rest of the paper is organized as follows. Section II provides a review of related work in the area of keyword and keyphrase identification. Section III presents the proposed approach, including the dataset, feature extraction, and classification techniques. Section IV discusses the results of the experiments, and Section V concludes the paper.

2.RELATED WORK

Keyword and keyphrase identification have been a significant research topic in the field of natural language processing and information retrieval. Various approaches have been proposed to extract keywords and keyphrases from textual data, including frequency-based methods, syntactic and semantic-based methods, and machine learning techniques. The keyword/ keyphrase extraction process involves five steps [1]. The text is first preprocessed, as in any other NLP task to remove unwanted symbols, images, url, etc. The next step is identification of candidate keyphrases which may be identified through n-grams, parts of speech tags like nouns, etc. This is followed by feature selection and selection of the keyphrases which can be done by using unsupervised or supervised approach. The final step is the performance evaluation of the model using either automatic or manual metrics or a combination of both. The actual execution of all steps largely depends on whether the approach used is unsupervised or supervised.

Unsupervised techniques for keyword/ keyphrase extraction

Unsupervised techniques work by ranking the identified candidates and finally extracting the top n candidates. There are several unsupervised

techniques for keyword and keyphrase extraction from text. Few of these are as follows:

Frequency-based methods

These methods rely on the frequency of occurrence of each word or phrase in the text. Words or phrases that appear more frequently are considered more important. Examples of such methods include TF-IDF (Term Frequency-Inverse Document Frequency) and RAKE (Rapid Automatic Keyword Extraction).

TF-IDF is a statistical technique utilized to determine the significance of a word in a particular document within a group of documents. It involves computing two metrics: the frequency of the word in the document, and the inverse document frequency of the word across the entire collection of documents. By evaluating these metrics, TF-IDF can provide insight into how important a given word is to a particular document. It is computes by the formula shown in Eq. 1

$$TF\_IDF = TF * IDF \quad (1)$$

where TF is computed as shown in Eq. 2

$$TF = \frac{No\ of\ times\ a\ term\ occurs\ in\ a\ document}{Total\ no\ of\ terms\ in\ the\ document} \quad (2)$$

and IDF is computed by Eq. 3

$$IDF = \log \frac{Total\ no\ of\ documents\ in\ the\ corpus}{No\ of\ documents\ in\ which\ the\ term\ occurs} \quad (3)$$

Although TF-IDF is easily computable, this method has the disadvantage that it does not take into account the context of the word/ phrase [2].

2. Graph-based methods

These methods represent the text as a graph and use graph-based algorithms to identify important nodes. Examples of such methods include TextRank [3] and LexRank [4]. By considering the significance of connected words and recursively computing the significance of each word within the graph, TextRank is Google's Pagerank based technique that identifies the importance of a word and selects the most highly ranked words as keywords [3]. A stochastic graph-based technique is employed by LexRank [5], a natural language processing approach, to determine the relative significance of textual units. Graph-based methods have the advantage of being domain independent but are able to extract keywords/ phrases only from one document at a time and is computationally expensive [2].

3. Clustering methods

These methods group together words or phrases that are similar to each other based on their context. Examples of such methods include K-

means and Hierarchical clustering. One algorithm that uses the clustering approach is TopicRank (TR) method [6]. To begin with, it conducts text preprocessing to extract potential phrases. These phrases are then sorted into different topics via hierarchical agglomerative clustering. In the following phase, a topic graph is created, and the edges between topics are assigned weights based on a metric that takes into account the offset positions of phrases within the text. Finally, TextRank is applied to rank the topics, and the most significant N topics (as determined by the ranking) have their initial keyphrase candidate selected.

4. Latent Semantic Analysis (LSA)

LSA is a mathematical technique that analyzes relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms. While the primary goal is to match relevant documents with keywords, the concept behind LSA is to compare the meanings or concepts of words rather than the words themselves. LSA is capable of analyzing the connection between a group of documents and the terms they contain by generating a collection of concepts that are related to both the documents and the terms [7].

5. Topic modeling

Topic modeling is a statistical modeling technique that identifies abstract topics that occur in a collection of documents. Examples of such methods include Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). Similar to LSA, LDA also involves verifying topic assignments for every word in each document, with the entire collection of documents being cycled through repeatedly [8]. This iterative updating is crucial to LDA's ability to produce a coherent set of topics as the final solution. In both LSA and LDA, documents are organized into clusters, with each cluster being given a set of keywords to define its meaning. The number of clusters to be created must be predetermined. LDA is limited has the ability to extract most general keywords.

B. Supervised techniques for keyword/ keyphrase extraction

Supervised techniques treat the problem as a classification problem and classifies the words/ phrases into being key or non-key. Initially, these techniques generate a training set that is labeled by developing characteristics for potential phrases or words in the text. By referring to the related gold-standard list, each phrase is identified as positive or negative. The resulting training set is then employed to create a predictive model that categorizes words (or phrases) in new documents as either a keyword or non-keyword [9]. Machine learning techniques that have been used include SVM [10], Naive Bayes [11], Random Forest, and neural networks [9]. SVM is a popular machine learning algorithm that works by identifying a hyperplane that separates the data into two

classes. Naive Bayes is a probabilistic algorithm that assumes that the features are independent of each other. Random Forest is an ensemble learning method that combines multiple decision trees to improve the accuracy. Neural networks are a class of machine learning algorithms that are inspired by the structure and function of the human brain.

Though some research has been done on keyword extraction from Indian languages [12] [13] [9] in the context of Indian domain specific text, there has been limited research on keyword and keyphrase identification.

In conclusion, the literature survey indicates that there have been several studies on keyword and keyphrase extraction from textual. While traditional methods like TF-IDF and LSA have been used, machine learning techniques such as SVM, Naive Bayes, and Random Forest, as well as deep learning techniques like CNNs, have shown promising results. Furthermore, social media data, political speeches and debates, and manifestos can be used as sources of textual data for keyword and keyphrase extraction in the Indian election domain. Hence, the objective of this work is to identify and extract keywords and phrases that are specific to the domain of Indian Elections.

3. METHODOLOGY

The methodology of keyword and phrase extraction is illustrated in Fig 1. It involves collection of data in various formats, preprocessing and cleaning the data to bring it into a format suitable for further processing, extraction using various models, ranking and selection of top n keywords and keyphrases. Since annotated dataset for the Indian Election domain are not available for training, unsupervised statistics based methods were used in this work. To improve the accuracy of the methodology to work for the Indian Election domain a Named Entity Recognition module was included since it is obvious that phrases like “Election Commission”, “Model Code of Conduct”, “voter id”, etc. are important keyphrases from the perspective of Indian Election domain.

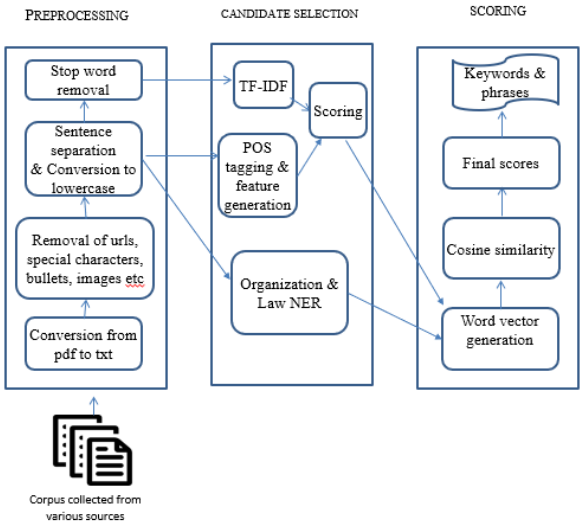
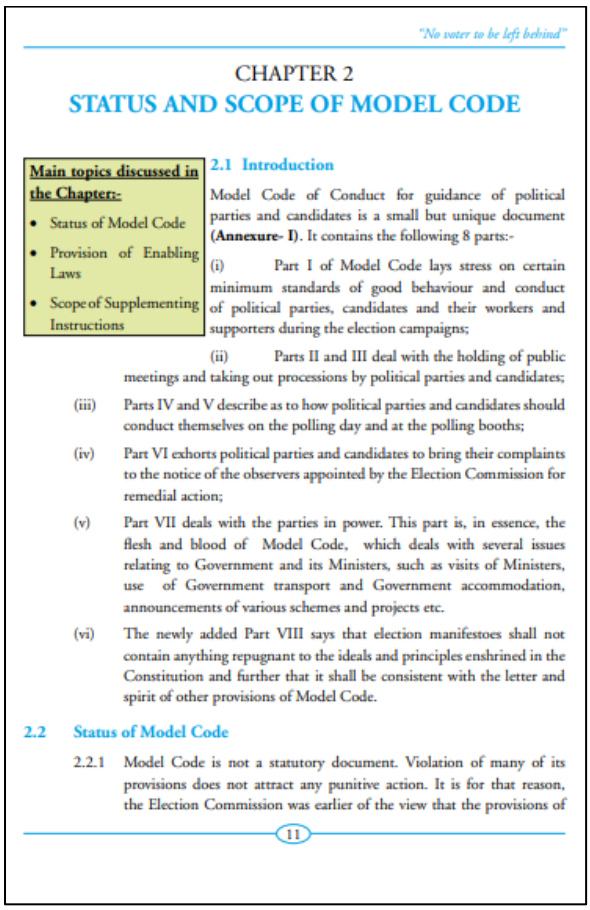


Fig 1: Methodology

A. Corpus Collection

The dataset for this work was collected from various resources and was in different formats. The majority of the dataset was downloaded from the website of Election Commission of India - https://eci.gov.in/. The website contains a large amount of textual data in pdf files in the form of manuals, compendiums, instructions, etc. The website also contains other formats too such as ppt and excel but for this work only pdf files and information available on the internet such as news items, blogs and tweets were used. Fig. 2 shows sample page from the Model Code of Conduct manual.



"No voter to be left behind"

CHAPTER 2

STATUS AND SCOPE OF MODEL CODE

Main topics discussed in the Chapter:-

- Status of Model Code
- Provision of Enabling Laws
- Scope of Supplementing Instructions

2.1 Introduction

Model Code of Conduct for guidance of political parties and candidates is a small but unique document (**Annexure- I**). It contains the following 8 parts:-

(i) Part I of Model Code lays stress on certain minimum standards of good behaviour and conduct of political parties, candidates and their workers and supporters during the election campaigns;

(ii) Parts II and III deal with the holding of public meetings and taking out processions by political parties and candidates;

(iii) Parts IV and V describe as to how political parties and candidates should conduct themselves on the polling day and at the polling booths;

(iv) Part VI exhorts political parties and candidates to bring their complaints to the notice of the observers appointed by the Election Commission for remedial action;

(v) Part VII deals with the parties in power. This part is, in essence, the flesh and blood of Model Code, which deals with several issues relating to Government and its Ministers, such as visits of Ministers, use of Government transport and Government accommodation, announcements of various schemes and projects etc.

(vi) The newly added Part VIII says that election manifestoes shall not contain anything repugnant to the ideals and principles enshrined in the Constitution and further that it shall be consistent with the letter and spirit of other provisions of Model Code.

2.2 Status of Model Code

2.2.1 Model Code is not a statutory document. Violation of many of its provisions does not attract any punitive action. It is for that reason, the Election Commission was earlier of the view that the provisions of

11

Fig. 2: A sample page of one document in the corpus

B. Preprocessing the corpus

It is evident from Fig. 2 that the collected corpus needed a lot of pre-processing and data cleaning to bring to a format suitable for applying any keyword/ keyphrase extraction technique. The process of data preprocessing is critical for converting raw data into valuable information. It is not advisable to directly feed unprocessed data into machine learning programs, as raw real-world data, including text, images, videos, and tables, is often disorganized and messy. Such unstructured data can cause errors and inconsistencies, and thus must be cleaned and analyzed beforehand. In data preprocessing, the text was converted to lowercase and any images, tables, whitespaces, special characters, URLs, stop words, itemized bullet and numbering, and numbers were removed. Additionally, stemming and lemmatization techniques were applied to extract the underlying meaning of words.

C. Named Entity Recognizer

Named Entity Recognition (NER) is a branch of Natural Language Processing (NLP) that uses automated techniques to identify and classify named entities in a given text. These entities can include various types of information such as people's names, organizations, locations, dates, times, and numerical values, among others. By automatically detecting and categorizing these entities, NER can help businesses and organizations derive valuable insights from large, unstructured datasets [14]. Named entity recognition was applied on the collected corpus to generate a list of keywords and phrases. Table I shows some of the generated named entities. After observing the entities, it was decided that the most relevant entity for keyword and phrases are the organization and law entities. For example, entity “Election Commission” is an organization and “Model code of conduct” is law. The other entities such as location, date or geo-political entities do not really add to the domain specific knowledge. Hence, a list of the organizational entities was prepared to increase the probability that these are identified as candidate keywords and phrases.

Table I: Type of Named Entities

SN	***Type of NE***	***Description/ Example***
1	Geo-political	Countries, States, Cities, Districts, etc.
2	Organization	E.g. Election Commission of India, the Ministry of Information and Broadcasting
3	Cardinal	Numbers (includes numbers written as words e.g. “six”
4	Date	Whole dates or part of dates e.g. April
5	Person	Name of a person
6	Law	Constitution, Model code of conduct, etc.

D. Candidate Identification

The candidate keywords were identified using TF_IDF since it has a robust nature. To select the keyphrases, phrases whose constituent parts had high scores and whose parts-of-speech tags had the form, JJ, NNP, NNS, NNP S, NN (J – Adjective, N – Noun, NP – Noun phrase, S - plural) were selected as these have been proven to give good results [11].

E. Generation of Keywords and Keyphrases

Each candidate's score is determined by multiplying the features together. Additionally, taking into consideration the fact that larger document datasets tend to have more keyphrases per document, the top-N candidates (where N is equal to 2.5 times the base-10 logarithm of the document size) are selected for each document. This value of 2.5 was used as it has been proven to work [15]. Lastly, word vectors of the candidates as well as named entities identified were generated. Pairwise cosine similarity between each vector of named entity words and other word vectors computed. Cosine similarity is a metric that assesses the similarity between two vectors in an inner product space by calculating the cosine of the angle between them. This metric indicates whether two vectors are pointing in a similar direction or not. Text analysis frequently employs cosine similarity to evaluate document similarity [16]. This cosine similarity was also multiplied to get the final score. The keywords and phrases with a score higher then threshold were then extracted.

4. RESULTS

Fig. 3's word cloud depicts some of the most important terms within the domain of the Indian Election Commission.

Fig. 3 Word cloud

Fig. 4 shows some of the identified leywords and keyphrases extracted by applying the methodology.

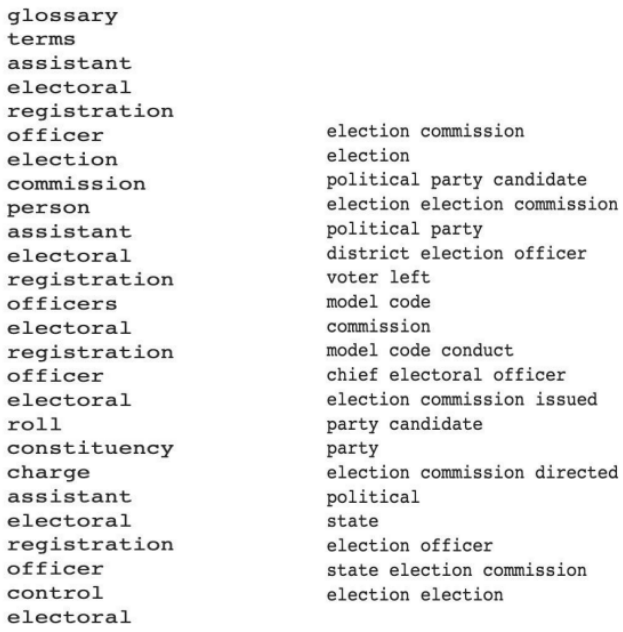


Fig. 4 Some identified keywords and keyphrases

Precision, recall, F1 score, and accuracy are the commonly used metrics to evaluate the performance of keyphrase extraction models. Precision measures the proportion of the extracted phrases that are actually relevant to the domain being studied [17]. It is calculated as shown in Eq. 4

$$Precision = \frac{No\ of\ correctly\ extracted\ phrases}{Total\ no\ of\ extracted\ phrases} \quad (4)$$

A high precision score indicates that the model is accurately identifying relevant phrases while minimizing false positives. Recall measures the proportion of relevant phrases that are correctly extracted by the model [17]. It is calculated as shown in Eq. 5.

$$Recall = \frac{No\ of\ correctly\ extracted\ phrases}{Total\ no\ of\ phrases\ in\ the\ corpus} \quad (5)$$

A high recall score indicates that the model is effectively identifying relevant phrases while minimizing false negatives. F1 score is the harmonic mean of precision and recall. It provides a single score that balances precision and recall [17]. It is calculated as shown in Eq. 6.

$$F - measure = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (6)$$

A high F1 score indicates that the model is performing well in both precision and recall, meaning that it is accurately identifying relevant phrases while minimizing false positives and negatives. Accuracy measures the proportion of all correctly identified phrases (both

relevant and irrelevant) out of the total number of phrases [14]. It is calculated as shown in Eq. 7.

$$Accuracy = \frac{No\ of\ correctly\ identified\ phrases}{Total\ no\ of\ phrases} \quad (7)$$

A high accuracy score indicates that the model is performing well in identifying both relevant and irrelevant phrases. These metrics are essential for evaluating the effectiveness of keyphrase extraction models as they provide valuable insights into the accuracy, completeness, and balance of the extracted phrases.

All these performance metrics requires a data set. The publicly available datasets are of three categories – news, scientific papers and abstracts of papers. As such we were unable to find a relevant data set for Indian Election domain due to which the above-mentioned performance metrics could not be evaluated. However, the precision metric was employed to evaluate the performance of the keyword extraction process. To compare the precision of different models, two human annotators independently evaluated the extracted keywords. An extracted keyword was considered correct only when both annotators agreed. The metrics was computed by selecting the highest ranked 5, 10,15 and 20 words considering TF-IDF with POS and using all three. Table II shows the comparison between these three. It can be seen from the table that the better precision is achieved using the domain specific knowledge

Table 2: Performance Evaluation

Approach	Precision@5	Precision@10	Precision@15	Precision@20
TF_IDF & POS	.389	.39	.395	.396
TF_IDF, POS & NER	.42	.423	.43	.43

CONCLUSION

This paper illustrated a methodology for keyword and keyphrase extraction, which involves collecting data in various formats, preprocessing and cleaning the data, extracting keywords and keyphrases using various models, ranking and selecting the top candidates. The methodology was applied to the Indian Election domain using unsupervised statistics-based methods due to the lack of annotated datasets. The inclusion of a Named Entity Recognition module was necessary to improve the accuracy of the methodology, and the results showed the importance of organizational and law entities in this

domain. The TF-IDF method was used for candidate identification, and the selected candidates were scored using word vectors and pairwise cosine similarity. The top keywords and keyphrases were extracted based on a threshold score. The results, depicted in word clouds and tables, show the effectiveness of the methodology in extracting relevant keywords and keyphrases in the Indian Election domain. Overall, this methodology can be applied to other domains and can help researchers and organizations in extracting valuable insights from large, unstructured datasets.

REFERENCES

1. F. Z. F. A. Z. E. h. B. Lahbib Ajallouda, "A Systematic Literature Review of Keyphrases," *International Journal of Interactive Mobile Technologies ,* vol. 16 , no. 16 , pp. 31-58, 2022.
2. M. T. A. Isabella Gagliardi, "Semantic Unsupervised Automatic Keyphrases Extraction by Integrating Word Embedding with Clustering Methods," *Multimodal Technologies and Interaction,* vol. 4, no. 2, 2020.
3. P. T. Rada Mihalcea, "TextRank: Bringing Order into Text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004.
4. A. Kumar, A. Sharma, S. Sharma and S. Kashyap, "Performance analysis of keyword extraction algorithms assessing extractive text summarization," in *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, Jaipur, India, 2017.
5. F. L. J. C.-C. Asahi Ushio, "Back to the Basics: A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, 2021.
6. F. B. B. D. Adrien Bougouin, "Keyphrase Annotation with Graph Co-Ranking," in *COLING 2016, 26th International Conference on Computational Linguistics*, Osaka, Japan,, 2016.
7. N. E. Evangelopoulos, "Latent semantic analysis," *Wiley Interdisciplinary Reviews: Cognitive Science,* vol. 4, no. 6, pp. 683-692, 2013.
8. X. Liu, Z. Zhang, B. Li and F. Zhang, "Keywords Extraction Method for Technological Demands of Small and Medium-Sized Enterprises Based on LDA,," in *2019 Chinese Automation Congress (CAC)*, Hangzhou, China, 2019.
9. V. B. Swagata Duaria, "Complex Network based Supervised Keyword Extractor," *Expert Systems with Applications,* vol. 140, no. 1, p. 112876, 2020.
10. C. Zhang, "Automatic keyword extraction from documents using conditional random fields," *Journal of Computational Information Systems ,* vol. 4, no. 3, pp. 1169-1180, 2008.
11. F. B. A. G. S. D. G. Cornelia Caragea, "Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach," in *Proceedings*

of the 2014 conference on empirical methods in natural language processing (EMNLP), Doha, Qatar, 2014.

12. V. g. Kamaldeep Kaur, "KEYWORD EXTRACTION FOR PUNJABI LAGUAGE," *Indian Journal of Computer Science and Engineering ,* vol. 2, no. 3, pp. 364-370, 2011.
13. S. Sifatullah and S. Aditi, "Keyword and keyphrase extraction from single Hindi document using statistical approach," in *Proceedings of 2nd International Conference on Signal Processing and Integrated Networks (SPIN)*, Noida, India, 2015.
14. D. Jurafsky, Speech & language processing, Pearson Education India,, 2000.
15. B. M. F. R. Miguel Won, "Automatic extraction of relevant keyphrases for the study of issue competition," in *Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019*, La Rochelle, France, 2019.
16. M. K. J. P. Jiawei Han, "Getting to Know Your Data," in *Data Mining*, Science direct, 2012, pp. 39-82.
17. G. T. Eirini Papagiannopoulou, "A review of keyphrase extraction," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 10, no. 2, p. e1339, 2020.