IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 28, 2020, accepted November 17, 2020, date of publication November 20, 2020,
date of current version December 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3039548

TOP-Rank: A Novel Unsupervised Approach for Topic Prediction Using Keyphrase Extraction for Urdu Documents

AHMAD AMIN[1], TOQIR A. RANA[1], NATASH ALI MIAN[2], MUHAMMAD WASEEM IQBAL[3], ABBAS KHALID[1], TAHIR ALYAS[4], (Member, IEEE), AND MOHAMMAD TUBISHAT[5]

[1]Department of Computer Science and IT, The University of Lahore, Lahore 54590, Pakistan
[2]School of Computer and Information Technology, Beaconhouse National University, Lahore 53700, Pakistan
[3]Department of Computer Science and IT, The Superior College (University Campus), Lahore 54600, Pakistan
[4]Department of Computer Science, Lahore Garrison University, Lahore 54810, Pakistan
[5]School of Technology and Computing, Asia Pacific University of Technology and Innovation, Kuala Lumpur 57000, Malaysia

Corresponding author: Toqir A. Rana (toqirr@gmail.com)

ABSTRACT In Natural Language Processing (NLP), topic modeling is the technique to extract abstract information from documents with huge amount of text. This abstract information leads towards the identification of the topics in the document. One way to retrieve topics from documents is keyphrase extraction. Keyphrases are a set of terms which represent high level description of a document. Different techniques of keyphrase extraction for topic prediction have been proposed for multiple languages i.e. English, Arabic, etc. However, this area needs to be explored for other languages e.g. Urdu. Therefore, in this paper, a novel unsupervised approach for topic prediction for Urdu language has been introduced which is able to extract more significant information from the documents. For this purpose, the proposed TOP-Rank system extracts keywords from the document and ranks them according to their position in a sentence. These keywords along with their ranking scores are utilized to generate keyphrases by applying syntactic rules to extracts more meaningful topics. These keyphrases are ranked according to the keywords scores and re-ranked with respect to their positions in the document. Finally, our proposed model identifies top-ranked keyphrases as topical significance and keyphrase with the highest score is selected as the topic of the document. Experiments are performed on two different datasets and performance of the proposed system is compared with existing state-of-the-art techniques. Results have shown that our proposed system outperforms existing techniques and holds the ability to produce more meaningful topics.

INDEX TERMS Topic extraction, top-rank, keyphrase extraction, topic prediction, Urdu positional ranking.

I. INTRODUCTION

In the last two decades, with the enhancement in the use of World Wide Web (WWW), several news forums such as news channels, reporters or column writers broadcast their daily news and articles on their websites. This evolution in online news and other electronic forums has provided numerous challenging tasks for researchers where they have to find useful information from trillions of unstructured data records. On the basis of the latest research in natural language processing (NLP) and statistics, researchers have developed several new techniques for extraction of valuable information from a collection of documents using hierarchical or probabilistic models called topic model [1]. The key benefit of topic modeling is to determine patterns among words or phrases and clustering documents which share similar patterns. In other words, topic modeling is a reproductive model for documents which identifies a simple probabilistic technique by which documents can be produced. Furthermore, topic modeling is a statistical method that facilitates for organizing, summarizing and understanding large amount of textual data [2].

Topic modeling and keyphrase extraction techniques assist to identify the title of the document which helps readers to choose most relevant documents with the help of the title. However, to assign a title for the document one has to read the whole document and then assign the most appropriate title.

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

This is a time consuming task and requires manual interaction by a human. Therefore, there is a need to build such a system which can automatically read the document and assign the most appropriate title. A lot of efforts have been performed for topic prediction for different languages. In different applications, topic modeling has been applied for detecting and retrieving information [3]. Moreover, topic modeling has been applied to countless fields including text clustering, document tagging, film genre identification, sentiment analysis, etc. [4]–[6]. There are several techniques available for topic modeling to identify or to extract latent information from a text document e.g. Latent Semantic Analysis (LSA) [7] Probabilistic Latent Semantic Analysis (PLSA) [8] and Latent Dirichlet Allocation (LDA) [9]. However, very limited work has been performed for Urdu language due to its complex structure [10]–[12]. Furthermore, topic modeling with keyphrase extraction has been studied for languages like English, Chines, Arabic, etc. but no such technique exists in the literature for Urdu language.

Urdu is a morphologically rich but resourced poor language [13]. It is Pakistan's national language and more than 170 million people all over the world use it for the communication[1]. It is a language enriched with grammar and has a wide range of derivations and inflections in a single word which makes it a difficult language to process. Since Urdu is new in the field of NLP and information retrieval (IR), very few amounts of research work has been performed on it. Many models and tools developed for other languages cannot operate with Urdu language because of the completely distinct language structure [14]. English language is written from left to right but Urdu writing script is written from right to left. Recognition of phrases in English language is simple as compared to Urdu because English language follows some standards i.e. space insertion, notion of capitalization, etc. However, in the Urdu language there is no standard for space insertion and notion of word capitalization. Hindi and Urdu are closer only to the speakers of both languages, but the writing style of both languages is distinct.

Urdu text classification problem has been studied by several researcher in recent years [15]–[20]. However, these approaches relies on the classification of the documents and did not consider the title prediction. Assigning a title to the document is different from classification problem. As it deals with the title prediction for a single document and each document may have different title on the basis of information provided in the document, even for the same domain. On the other hand, documents classification categorized similar documents in one class, based on the similarity among text in the document and assumes that all the documents belong to the same domain. This is why assigning a title to the document becomes more challenging.

Building on above explanation, in this paper we have proposed a novel TOP-Rank approach for topic prediction by extracting top-ranked keyphrases in Urdu language. In the first step, the proposed approach pre-processes the text to remove invalid characters, stop words and sentence boundary identification. After preprocessing, our system identifies keywords and assigns rank (score) to each extracted keyword on the basis of its position in a sentence. After ranking keywords, proposed system extracts keyphrases of different sizes from the document based on the extracted keywords and these keyphrases are ranked by adding score of each keyword in the phrase. Once keyphrases are ranked, top-ranked keyphrases are selected and re-ranked by re-visiting the document and score of these keyphrases is updated based upon their occurrence in the document. Finally, keyphrase with the highest score is assigned as topic of the document. We have conducted experiments on Urdu language datasets which contains multiple documents from several different domains. The effectiveness of our proposed model is evaluated on these datasets and compared with the topic modeling-based state-of-the-art approaches. Experimental results have shown that our proposed model outperformed topic modeling-based approaches and have shown promising results on Urdu language dataset.

The rest of the paper is organized as follows: section two highlights related work and section three presents the proposed methodology. Explanation of dataset and experimental evaluations is given in section four and finally we conclude our work in section five.

II. RELATED WORK

Several supervised and unsupervised approaches have been developed for the topic modeling for different languages. In supervised line of research, classifiers are trained on the textual data and annotated with keyphrases that determine whether the document or phrase is a topical keyphrase or not. Huang *et al.* [21] proposed a supervised topic modeling technique Siamese Labeled Topic Model (SLTM) for English language at sentence level. The working mechanism was similar to the pLSA in which they distributed words on the basis of their labels and used artificial neural networks for training perspective. Wang *et al.* [22] proposed hierarchical Dirichlet process-based inverse regression (HDP-IR) model for the evaluation of e-commerce reviews. HDP-IR contained three components - non-parametric component, inverse regression and coupling component. The first component was used to build HDP to capture the uncertainty of data concerned with topics. Second component influenced by multinomial inverse regression (MNIR) model, while third component combined the first two components and integrated the topics into the logistic regression within MNIR model. Zeng *et al.* [23] proposed expectation-maximization algorithm for topic modeling by computing maximum likelihood. For distribution of the topics from document, they demonstrated the fast online expectation maximization (FOEM) which was able to converge LDA's probability function at local stationary point. By dynamically scheduling fast speed and streaming parameters for low memory use, FOEM was more efficient for lifelong topic modeling for big amounts of data. Li *et al.* [24]

[1]https://en.wikipedia.org/wiki/Urdu

designed generative modeling for multi-labeled classification of documents and trained two different extensions of LDA named as frequency LDA (FLDA) and dependency frequency LDA (DFLDA) for multi-label documents categorization task. These two models aimed to incorporate observations of labeled frequency and labeled dependency into the traditional LDA. FLDA used label frequency information to generate labeled Dirichlet prior to each document, while DFLDA introduced a topic-layer to capture co-occurrence relationships among labels.

In unsupervised approaches, various measures such as TF-IDF and topic proportions are used to identify topic associated terms. In topic prediction, keywords are ranked based on their relevance to the topic [25]. Latent semantic analysis (LSA) was developed for analyzing the relationships between a set of documents and the terms contained. Documents are compared by taking the angle cosine such that values closed to 1 represents documents are similar, whereas values closed to 0 represents different documents [7]. Bastani *et al.* [26] developed an intelligent system to analyze consumer complaints and labeled each document with different keywords and trained LDA model for topic prediction from these complaints. Venkatesaramani *et al.* [27] proposed two-step approach for topic modeling for short text collected from tweets and YouTube comments. They used TF-IDF based clustering to find similarity between comments. Zhang & He [28] proposed an approach for extracting topics for events on social media using reinforcement knowledge. Their methodology consisted of three steps: first they run their topic model based on word embedding and the structure of the conversation for mining preceding topic of each event. In the second step, they mined set of reinforced knowledge from previously extracted topic. Finally by using the reinforced knowledge sets they extracted the final topic for every event. Wang *et al.* [29] proposed a system for news-topic recommendation by extracting keywords from news articles. They proposed rapid automatic keyword extraction (RAKE) method for extracting keywords from online news and then ranking these extracted keywords by using position rank algorithm on the basis of syntactic rules. Alhawarat & Hegazi [30] proposed topic modeling technique for Arabic language news data set. They used LDA and k-means clustering algorithms. For topic prediction they reduced vector space model and then extracted the hidden topic from the document as a feature selector.

Several approaches have adopted keyphrase extraction method. Bougouin *et al.* [31] proposed an approach for keyphrase extraction for topic prediction and used a topic-based model with clustering. After making the clusters of topics, a graph was generated where topic clusters were considered as vertices and edges created by calculating keyphrases in the clusters that appeared together. Distance was calculated between clusters and this distance created edges between two clusters. They used text rank algorithm to rank their topics and finally keyphrase having higher rank considered as topic. Danilevsky *et al.* [32] did not consider the length of the keyphrase and identified topics and grouped them in clusters. Each cluster is assigned words from the document and these words are formed into keyphrases and ranked based on purity, completeness, and coverage and finally, the topmost keyphrases were selected as topic. Parveen *et al.* [33] proposed a graph-based model where they considered every node as a topic. For each edge in the graph, they normalized according to their length so that long sentences do not get any benefit. Furthermore, they utilized Hyperlink-induced topic search (HITS) algorithm to rank sentences. Boudin [34] extracted keyphrases by using multipartite graphs and sentence clusters to incorporate the topical knowledge, and Alfarra and Alfarra [35] proposed a graph-based technique for extracting keyphrases in a single document which utilized phrases and terms in a sentence rather than focusing structure of the document.

Wan *et al.* [36] proposed an unsupervised graph-based approach for both summarization and keyword extraction. They generated sentence-to-sentence graph, sentence-to-word graph and word-to-word graph. Another graph based approach was proposed by Danesh *et al.* [37] which used TF-IDF, term length and position of first occurrence (PFO) of keywords for ranking mechanism. The main contribution was to use PFO which ranked keyphrases passes on a certain threshold and decreases the term frequency score if it appears in other terms. Ali, Wang and Haddad [38] assigned each word a syntactic category (i.e. noun, adjective, etc.) and identified several syntactic patterns, and phrases were extracted according to the syntactic pattern. Furthermore, these keyphrases were ranked using TF-IDF and text rank algorithm. Corina and Cornelia [39] developed position rank graph-based approach for keyphrase extraction. They build graphs where words represent nodes and edges represent weights. The edges were assigned weights based on how many times these words appear together.

Shakeel *et al.* [10] proposed a topic modeling technique for Urdu language based upon standard LDA as Urdu-LDA (ULDA). They utilized Gibbs sampling technique along with Markov Chain Monte Carlo (MCMC) algorithm for extracting topics from Urdu text document. Rehman *et al.* [11] proposed probabilistic topic model with statistical variational Bayes approach for Urdu documents (VB-ULDA). They described two versions of VB-ULDA, with stemmer VB-ULDA (WS) and without stemmer VB-ULDA (WiS) for topic modeling of Urdu text article. Rehman *et al.* [12] proposed non parametric Bayesian Hierarchical (hLDA) model for topic modeling in Urdu text articles (uhLDA). For statistical and probabilistic inference, they used Gibbs Sampling algorithm and extracted topics based on the hierarchies of the terms used in documents.

There exist several approaches for document classification in Urdu language. Ahmed *et al.* [16] proposed a SVM based classifier for Urdu news headline classification. They used predefined classes to group similar headlines in a single class. Zia, Akhtar and Abbas [17] presented a comparative study of different classification algorithms on Urdu

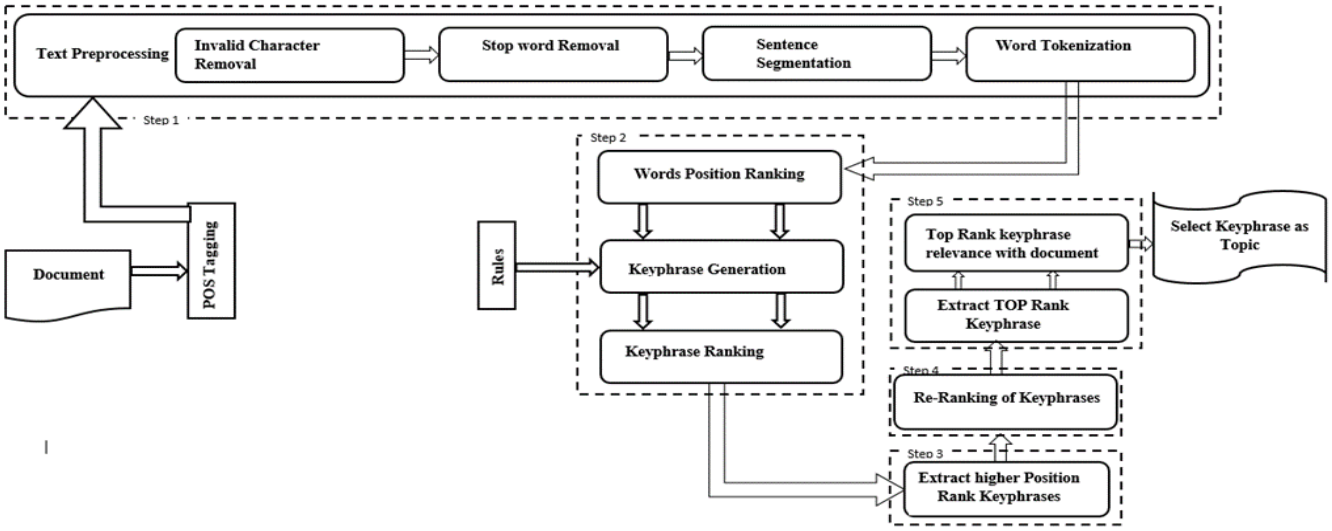

FIGURE 1. Proposed methodology.

document classification. They analyzed the classification algorithms with respect to the selected features to classify text. Akhtar *et al.* [19] used Single-layer Multisize Filters Convolutional Neural Network (SMFCNN) to classify Urdu documents and presented a comparison with several machine learning algorithms. Akhtar *et al.* [20] presented another comparison of deep learning algorithms where they selected four deep learning algorithms and compared their performance over Urdu document classification along with four machine learning algorithms. Rasheed, Banka and Khan [40] proposed a feature selection approach for Urdu news articles classification. They used LSI model to extract useful features from the news articles and used SVM for the classification.

III. PROPOSED MODEL

This section provides an overview of the proposed methodology for topic prediction using positioned-based top-ranked keyphrase extraction. Figure 1 illustrates the architecture of the proposed model which is divided into several steps: first step is the preprocessing of text data, second step is to extract positions of each word from the target document, generating keyphrases with the help of syntactic rules and ranking these keyphrases. Third step performs extraction of higher position rank keyphrases. To extract the most relevant and important keyphrases from an article, proposed system re-ranks keyphrases in step four and extracts top-rank keyphrases in step five and selects topic for the target document.

A. TEXT PREPROCESSING

Text pre-processing is an important step in NLP tasks as it transforms text into a more digestible form to achieve better performance of the algorithms. To achieve this, part-of-speech (POS) tagging is performed to allow syntactic filters because it helps for extraction of nouns, adjectives or pronoun phrases from the text document. For this purpose, we use well known POS tagger for Urdu language CLE[2]. After generating POS tags, several steps to clean the dataset have been performed e.g. removal of invalid characters, stop word removal and sentence segmentation as explained in upcoming subsections.

1) INVALID CHRACTER REMOVAL

In pre-processing, the first step is to remove invalid symbols like punctuation marks, links and special characters like !, ?, @, /, #, $, %, ^, &, ∗, (,), etc. from the text as these are unnecessary elements and hence it is better to remove all these invalid characters.

2) STOP WORD REMOVAL

In Urdu language, there are different kind of stop words known ''حروف جار'' (Haroof-e-Jaar or post position) similar to prepositions in the English language which appears before the object. Table 1 elaborates the difference between preposition in English and postposition in Urdu language.

TABLE 1. Postposition in urdu language.

The book is on the table.	کتاب میز پر ہے۔
Boy is writing on the paper.	لڑکا کاغذ پر لکھ رہا ہے۔

In English, on (''پر'') comes before the object table (''میز'') but in Urdu ''پر'' (on) comes after the object ''میز'' (table).

Table 2 shows some Urdu Haroof-e-Jaar (Postposition). These words provide no meaningful information about the document and therefore must be removed from the target documents.

[2]www.tech.cle.org.pk/services/text/pos

TABLE 2. Haroof-E-Jaar.

Urdu	Translation
میں	**Main**
نے	**Nay**
سے	**Say**
ہے	**Hay**
ہیں	**Hain**

TABLE 3. Haroof-E-Izafat.

Urdu	Translation
کا	Kaa
کے	Kay
کی	Kee
کو	Ko
اور	Or

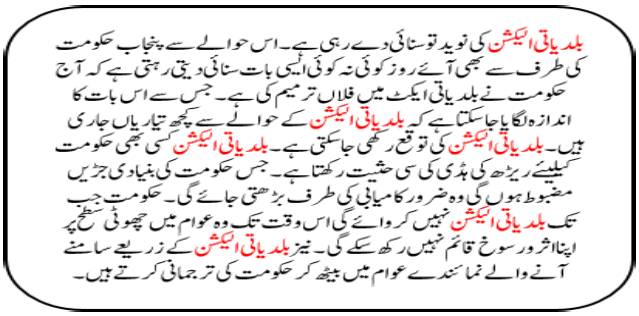
بلدیاتی الیکشن کی نوید توسنائی دے رہی ہے۔اس حوالے سے پنجاب حکومت کی طرف سے بھی آئے روز کوئی نہ کوئی ایسی بات سنائی دیتی رہتی ہے کہ آج حکومت نے بلدیاتی ایکٹ میں فلاں ترمیم کی ہے۔ جس سے اس بات کا اندازہ لگا یا جاسکتا ہے کہ بلدیاتی الیکشن کے حوالے سے کچھ تیاریاں جاری ہیں۔بلدیاتی الیکشن کی توقع رکھی جاسکتی ہے۔بلدیاتی الیکشن کسی بھی حکومت کیلیئے ریڑھ کی ہڈی کی سی حیثیت رکھتا ہے۔ جس حکومت کی بنیادی جڑیں مضبوط ہوں گی وہ ضرور کامیابی کی طرف بڑھتی جائے گی۔ حکومت جب تک بلدیاتی الیکشن نہیں کروائے گی اس وقت تک وہ عوام میں چھوٹی سطح پر اپنا اثر ورسوخ قائم نہیں رکھ سکے گی۔ نیز بلدیاتی الیکشن کے زریعے سامنے آنے والے نمائندے عوام میں بیٹھ کر حکومت کی ترجمانی کرتے ہیں۔

FIGURE 2. Keyphrase positioning in the document.

TABLE 4. Explanation about Haroof-E-Izafat.

Urdu	Translation
پاکستان کے وزیراعظم	Prime Minister of Pakistan
کھیلوں کے میدان	Play grounds
سائنس کی ایجادات	The inventions of science
کرکٹ کا میچ	The match of cricket
کشمیرکو بچاو	Save to Kashmir
چین اور پاکستان کی دوستی	The friendship of China and Pakistan
بابری مسجد کی شہادت	The demolition of Babri Mosque

Other than ''حروف جار'' (Haroof-e-Jaar), there are also some words that are used to link two words in Urdu and known as ''حروف اضافت'' (Haroof-e-Izafat) as in Table 3.

Haroof-e-Izafat are the words which creates relationship between nouns and adjectives. Table 4 illustrate examples for the use of Haroof-e-Izafat within a sentence. This can be observed that Hroof-e-Izafat are adding more explanation to understand meanings of the phrases. Therefore, proposed technique will remove all the stop words (Haroof-e-Jaar) and does not remove Haroof-e-Izafat as they express information useful for the identification of topics.

3) SENTENCE SEGMENTATION

In any language, a collection of words makes sentences and sentences create the document. Each sentence in a document contains significant information. To extract this information, the proposed system splits sentences on the basis of dash ''-'' having <SM> tag in the POS tagger for Urdu language. The position of each keyword will be counted from the start of each sentence because important keywords that contain significant information about the document occur on the starting positions of each sentence. Figure 2 highlights the impact of keywords positioning in the sentence. Although keyword ''حکومت'' is also frequent but ''بلدیاتی الیکشن'' has more importance in the document due to its position in a sentence and represent more meaningful topic.

B. WORD POSITION RANKING

After completing preprocessing, position of each keyword in a document is extracted by using proposed technique. The proposed system aggregates information from all positions of word occurrences in the document. The main idea behind the position ranking is to assign higher weight (or probability) to those words which are at the start of a sentence. For example, if a word appeared at 2nd, 5th, and 10th position in a document then its weight will be calculated as $1/2 + 1/5 + 1/10 = 0.8$. Equation 1 shows how the weight of each word w is calculated, where w_i is the particular word in a document, j is the position of the word. Summing up the positional weights for a given word aims to grant more confidence for frequently occurring words by taking into account the position weight of each occurrence.

$$w_i = \sum_{j=pos}^{w_i} \frac{1}{j} \tag{1}$$

However, it is to be noted that weight will remain same at every position in a document once it is calculated. For example, if the aggregated weight of the word ''پاکستان'' (Pakistan) is calculated as 1.6 in the target document then its aggregated weight will remain the same at every position wherever word ''پاکستان'' (Pakistan) occurs.

C. KEYPHRASE GENERATION

Keyphrases describe the most important topics about document, therefore to identify keyphrases, nouns and adjectives are considered as candidate words because they hold key information in a sentence for the identification of a topic. These words combined can represent more valuable information than their independent usage. The proposed system generates keyphrases by following some syntactic rules. These rules are:

R1: Each keyphrase will be comprised of a given size

R2: Each keyphrase will never start and end with Haroof-e-Izafat, e.g. ''کی'',''کے'',''کا'' and ''کو''

R3: After the first noun or adjective, if Haroof-e-Izafat is associated in the sentence then it will be included in the keyphrase

Algorithm 1 highlights the process for the generation of keyphrases. Algorithm takes POS tagged file as input and preprocesses it. This includes invalid character removing, stop word removal and sentence segmentation. First of all, algorithm ensures that keyphrase starts or ends with noun or adjective and Haroof-e-Izafat (''کا'',''کے'',''کی'',''کو'') do not appear at the start or end of any keyphrase. Line 10-11 will check that the start or end of the keyphrase and ensure that there should not be Hroof-e-Izafat (according to rule R2) and if any appears at these positions then such keyphrases will be discarded. The proposed system will never generate keyphrases like ''کے پاکستان کے'' because it does not produce any meaningful information. If the system is in the start or end of the keyphrase then line 12-18 generate keyphrases by joining nouns and adjectives. Lines 21-26 generate keyphrases if the system is not at the end or start of the keyphrase. Here rule R3 is applied and if any Hroof-e-Iazafat appears after the first noun of adjective of the keyphrase then it is considered as the part of the keyphrase. The algorithm generates several keyphrases from each sentence. The algorithm repeats itself for each sentence and generates the list of keyphrases as an output. At the end all keyphrases are extracted which start and end with a noun or adjective like ''پاکستان کے وزیراعظم''.

D. KEYPHRASE RANKING

Figure 3 illustrates the mechanism for keyphrase ranking where each keyword contains aggregated weight (calculated in subsection B) based upon its position in the document i.e. keyword ''پاکستان'' has the positional weight 1.6. Similarly keyword ''کے'' have weight 0.2 and ''وزیراعظم'' has weight 1.3. Keyphrase ''پاکستان کے وزیراعظم'' (The Prime Minister of Pakistan) has been ranked by adding the weight of each keyword which is 3.1.

Algorithm 1 Keyphrase Generation

```
1: Input: Tagged_Document
2: D = { }                    // list of keyphrases initially empty
3: keyphrase = Null           // String type variable for storing a
                                 single keyphrase
4: window_size = n            // Size of keyphrase
5: phrase_position = 1
6: Preprocessed_document ←
        Preprocess (Tagged_Document)
7:   for each Sentence in Preprocessed_document do
8:     for each Word in a Sentence do
9:       if phrase_position == 1 OR phrase_position ==
                                 window_size
10:          if word is Haroof-e-Izafat then
11:            break
12:          else if word is tagged noun OR adjective then
13:            keyphrase = keyphrase + word
14:            phrase_position + = 1
15:            if length(keyphrase) == window_size then
16:              D←keyphrase
17:              keyphrase←Null
18:            end if
19:          end if
20:
21:      else if word is noun OR adjective OR
                                 Haroof-e-Izafat then
22:            keyphrase = keyphrase + word
23:            phrase_position + = 1
24:            if length(keyphrase) == window_size then
25:              D←keyphrase
26:              keyphrase←Null
27:            end if
28:        end if
29:      end for
30:   end for
31: Output: Set of keyphrases D
```

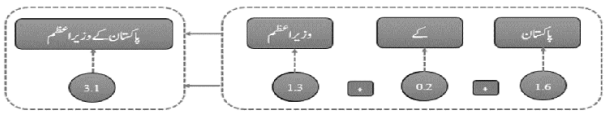

FIGURE 3. Keyphrase Ranking Mechanism.

The complete steps of keyphrase ranking are shown in Algorithm 2. This algorithm takes input the list of keyphrases generated by Algorithm 1 and score of keywords calculated in subsection B. The algorithm calculates the keyphrase rank by adding positional weight of each word and generates a list of ranking of all keyphrases as output. Line 5-7 add score of each word in a keyphrase to generate an overall score of the keyphrases.

E. EXTRACTING HIGHER RANK KEYPHRASES AND RE-RANKING

Keyphrases are comprised up with multiple words. To extract higher rank keyphrases from the target document, the

Algorithm 2 Keyphrase Ranking

```
Input: Set of keyphrases D and keywords scores W
R = { }                    // list for keyphrases ranks
keyphrase_rank = 0
for each keyphrase in D
    for each word in keyphrase
        keyphrase_rank + = W[word]
    end for
  R←keyphrase_rank
  keyphrase_rank = 0
end for
Output: Set of keyphrases ranks R
```

proposed system sorts keyphrases on the basis of their scores, extracts 10 higher rank keyphrases and further uses them for re-ranking.

After extracting higher rank keyphrases, next task is to extract most relevant keyphrase from the document. For this purpose, we have defined a novel re-ranking mechanism which revisits the document with keyphrase perspective and re-rank keyphrases based upon their occurrence in the document. For example, a keyphrase "پاکستان کے وزیراعظم" has initial rank 3.1 however on revisiting document if this keyphrase has been identified more than once then it means this keyphrase contains more abstract information about the document. Therefore, there is a need to re-rank this keyphrase by increasing its score based on its occurrence in the document and this helps to extract more relevant keyphrases from the document. The proposed model revisits the document for all top ten keyphrases. If any keyphrase appears more than one in the document then the rank of the keyphrase is incremented by one for each occurrence. By doing this the system will obtain top-rank keyphrases which are more accurate predicted topics of the target document. Algorithm 3 highlights the mechanism to re-rank keyphrases. If a keyphrase appears more than once in the document then against each occurrence its score is incremented by one.

IV. EXPERIMENTAL EVALUATION

This section presents the explanation of the experimental evaluation for the proposed model and elaborates our results including a comparison with related techniques.

Algorithm 3 Re-Rank Keyphrases

```
Input: higher_rank_keyphrases, document
count = 0
for each keyphrase in higher_rank_keyphrases
    for each occurrence of keyphrase in document
        count + = 1
    end for
    keyphrase_rank + = count
  count = 0
end for
Output: Keyphrase_rank
```

A. DATASET AND EVALUATION METRICS

Evaluation measures are a vital part to assess the performance of the building model. Almost all evaluation measures depend on the nature of the data. To calculate the performance of the proposed system, a number of experiments are performed on a variety of datasets. We have choose two datasets, first dataset (D1) is prepared by us which contain 640 documents. The dataset has been collected from different websites and news articles e.g. express.pk/, bbc.com/urdu/, urdupoint.com/ and urdu.geo.tv etc. Dataset contains documents from five different domains which include politics, sports, entertainment, health and economy. Each domain consists of more than 120 different documents and each document contains several sentences. Table 5 presents detailed overview of the dataset D1. Second dataset (D2), Northwestern Polytechnical University Urdu (NPUU), was prepared by [19] and contains more than 10,000 documents from six different domains which include business, crime, entertainment, politics, science and technology and sports. Table 6 presents detailed overview of the dataset. D2 dataset contains documents

TABLE 5. Summary of dataset D1.

Domain	Documents		
	No. Documents	Percent	Words
Politics	120	19%	45,704
Sports	125	19%	43,717
Entertainment	140	22%	57,840
Health	135	21%	51,318
Economy	120	19%	41,144
Total	**640**		**2,39,723**

TABLE 6. Summary of dataset D2.

Domain	Documents		
	No. Documents	Percent	Words
Business	3,328	31%	1,125,700
Crime	847	8%	210,021
Entertainment	1,913	18%	549,117
Politics	1,636	15%	573,786
Science & Tech	1,655	15%	779,331
Sports	1,440	13%	373,801
Total	**10,819**		**3,611,756**

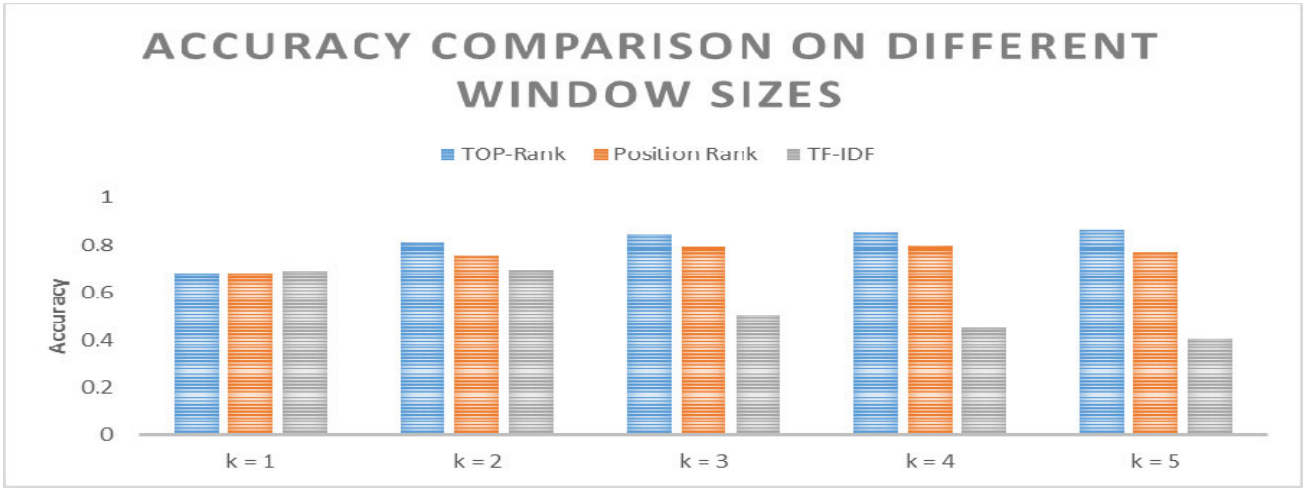

FIGURE 4. Accuracy comparison of TOP-Rank over dataset D1.

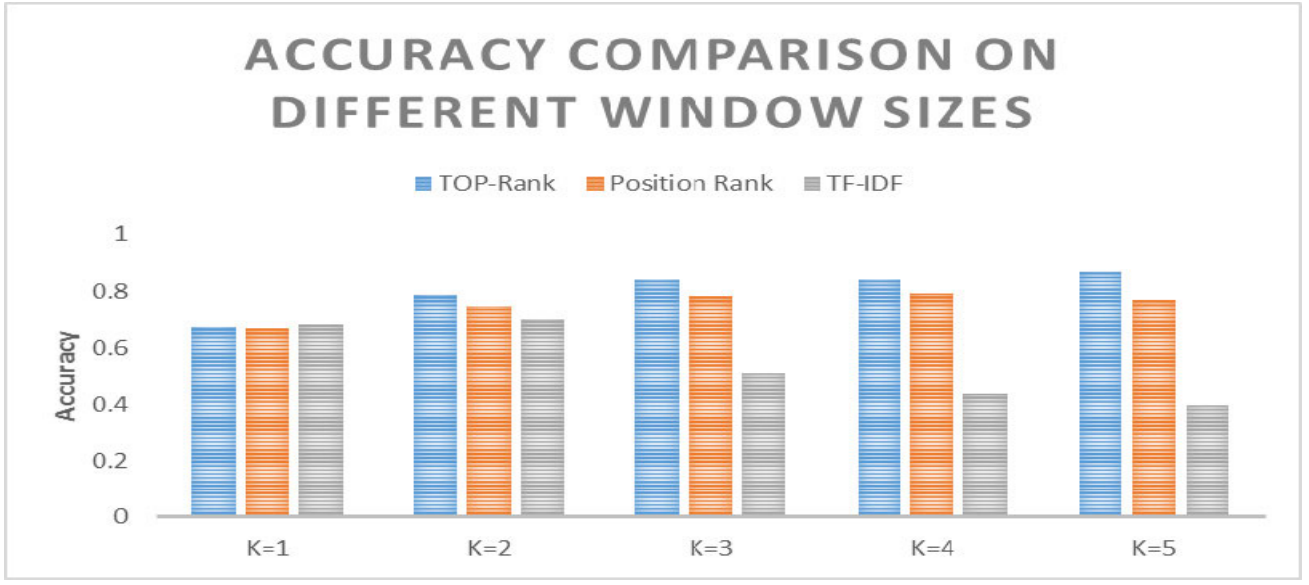

FIGURE 5. Accuracy comparison of TOP-Rank over dataset D2.

collected from news websites and annotated manually by human annotators. Table 5 and 6 present the domains of documents along with total number of documents in each domain, percentage of each domain in the dataset and total number of words in all documents in the domain for both D1 and D2 datasets respectively.

B. RESULT ANALYSIS

To evaluate the proposed system, a documents from both datasets D1 and D2 were executed multiple times to extract top-rank keyphrase as topics for documents. Accuracy of the proposed system is measured by following equation.

$$Accuracy = \frac{no.\ of\ true\ predicted\ documents}{total\ no.\ of\ documents} \times 100 \quad (2)$$

The topics predicted by the system are evaluated by TOP-Rank (proposed model), position rank and TF-IDF predictions on different window sizes (keyphrase length). To check the correctness of the proposed system, systematic procedure marked the true predicted topic if it fulfills the condition that topical keyphrase should be on the top out of the top ten predicted keyphrases for all three mechanisms TOP-Rank, position rank and TF-IDF.

For result analysis, keyphrases with different sizes are selected which are (k = 1) unigram, (k = 2) bigram, (k =3) trigram, and (k = 4 and k = 5) considered as n-gram. Unigram contains single word e.g. ''پاکستان'' (Pakistan), bigram contains two words e.g. ''پاکستان وزیراعظم'' (Prime Minister Pakistan), trigram contains three words e.g. ''پاکستان کے وزیراعظم'' (Prime Minister of Pakistan) and

TABLE 7. Results comparison of top-rank keyphrase relevance with position rank and TF-IDF over D1.

Domain	Technique	k = 1	k = 2	k = 3	k = 4	k = 5
Politics	TOP-Rank	0.725	**0.808**	**0.883**	0.867	**0.858**
	Position Rank	**0.733**	0.775	0.808	**0.892**	0.834
	TF-IDF	0.708	0.651	0.441	0.408	0.358
Sports	TOP-Rank	0.684	**0.811**	**0.846**	**0.853**	**0.861**
	Position Rank	0.679	0.752	0.782	0.801	0.757
	TF-IDF	**0.696**	0.692	0.479	0.432	0.384
Entertainment	TOP-Rank	0.701	**0.807**	0.814	**0.887**	**0.828**
	Position Rank	0.685	0.785	**0.836**	0.737	0.721
	TF-IDF	**0.717**	0.671	0.464	0.436	0.429
Health	TOP-Rank	0.683	**0.821**	0.838	**0.842**	**0.882**
	Position Rank	0.678	0.752	**0.859**	0.816	0.821
	TF-IDF	**0.691**	0.714	0.625	0.532	0.502
Economy	TOP-Rank	0.625	**0.817**	**0.842**	**0.825**	**0.891**
	Position Rank	0.617	0.692	0.701	0.758	0.717
	TF-IDF	**0.645**	0.742	0.533	0.451	0.367

TABLE 8. Results comparison of top-rank keyphrase relevance with position rank and TF-IDF over D2.

Domain	**Technique**	**K = 1**	**K = 2**	**K = 3**	**K = 4**	**K = 5**
Business	TOP-Rank	0.624	**0.816**	**0.841**	**0.824**	**0.891**
	Position Rank	0.617	0.692	0.700	0.758	0.716
	TF/IDF	**0.651**	0.741	0.532	0.449	0.366
Crime	TOP-Rank	0.628	**0.809**	**0.828**	**0.832**	**0.864**
	Position Rank	0.619	0.702	0.711	0.756	0.762
	TF/IDF	**0.652**	0.743	0.534	0.448	0.364
Entertainment	TOP-Rank	0.698	**0.811**	0.821	**0.856**	**0.835**
	Position Rank	0.688	0.789	**0.833**	0.735	0.718
	TF/IDF	**0.708**	0.672	0.465	0.436	0.427
Politics	TOP-Rank	0.727	**0.807**	**0.881**	0.864	**0.856**
	Position Rank	**0.736**	0.774	0.805	**0.888**	0.831
	TF/IDF	0.707	0.648	0.440	0.408	0.356
Science & tech	TOP-Rank	0.685	**0.822**	0.836	**0.844**	**0.888**
	Position Rank	0.676	0.754	**0.857**	0.812	0.821
	TF/IDF	**0.688**	0.715	0.618	0.534	0.500
Sports	TOP-Rank	0.687	0.665	**0.827**	**0.838**	**0.863**
	Position Rank	0.673	**0.756**	0.784	0.810	0.753
	TF/IDF	**0.695**	0.684	0.469	0.345	0.381

keyphrases contains four or five words are considered as n-gram e.g. "پاکستان کے وزیراعظم عمران" (Prime Minister of Pakistan Imran), "پاکستان کے وزیراعظم عمران خان" (Prime Minister of Pakistan Imran khan) respectively. Keyphrases with different window sizes (k) are compared with higher positional rank keyphrase and TF-IDF. The averaged keyphrase relevance with document is given in Table 7 and 8 for each window size, i.e. unigram, bigram, trigram and n-gram for datasets D1 and D2 respectively. Each document contains an average forty original assigned keyphrases but the system will extract only the top ten higher ranked phrases.

Table 7 and 8 represent average top-rank keyphrase relevance with target document for all classes at each window size i.e. unigram, bigram, trigram and n-gram over dataset

D1 and D2 respectively. It can be noted that TF-IDF performs better in some cases for unigram due to the single keyword extraction. However, its accuracy starts decreasing when size of keyphrases increases and produces more inaccurate results as compared to position rank and TOP-Rank keyphrases. The main reason is that it extracts keyphrases only based on the frequency of the words instead of its positional influence in the document. Position rank produces lower accuracy as compared to TOP-Rank because its contains only positional weights of the words in keyphrases but in some cases positional rank keyphrases performed better than TOP-Rank on different window sizes in different classes due to the limitation of re-ranking mechanism. Because in TOP-Rank, we re-ranked keyphrases for some cases which resulted in some irrelevant keyphrases to become top-rank which is considered as false predicted topic for TOP-Rank mechanism. But for overall evaluation, it can be noted that TOP-Rank outperformed both positional rank keyphrase and TF-IDF.

We can observe from Table 7 and 8 that against unigram TOP-Rank, position rank and TF-IDF produced similar results but as a whole TF-IDF performs better at unigram. Position rank performs better at trigram in politics and health in D1 and entertainment and simlarly it perfroms better for science and technology and entainment in NPUU. After increasing the window size for keyphrase prediction, it is clearly indicated that TOP-Rank performs better as compared to the both datasets and produces more accurate and meaningful topics. Moreover, it is to be noted that more accurate results were found on tri-gram and n-gram because when the window size was increased, keyphrases provide more accurate information about the document while single

keywords i.e. unigram cannot deliver the accurate information about the document. Further, it is evident that TOP-Rank performs better as compared to position rank and TF-IDF resultantly producing more accurate topics for documents in both the datasets.

Figure 4 and 5 present an overall comparison of the proposed TOP-Rank model with position rank and TF-IDF for both D1 and D2 datasets respectively. This is clear from the diagram that TF-IDF has the highest accuracy for unigram however, position rank and TOP-Rank have also produced comparable performance. On the other hand, as we increase the length of the keyphrases, our proposed model TOP-Rank shows highest accuracy.

These figures clearly depicts that on unigram all three yield similar results but when increasing window size at trigram and n-gram, TOP-Rank outperforms the other two and extracts more accurate results. From the diagrams it is evident that TOP-Rank outperformed both position rank and TF-IDF and produced better results. This establishes the effectiveness of the re-ranking mechanism adopted in TOP-Rank model.

V. CONCLUSION

Topic modeling is the key method in machine learning and NLP to extract significant information from a document. This information helps to identify important topics from the document which further could be utilized for topic prediction. Topic modeling has been widely explored in different languages however, there is limited work for topic modeling for Urdu documents. In this research we have introduced a framework to extracts meaningful topics from Urdu documents by using TOP-Rank keyphrase extraction based on their positional weight. There exists no technique on keyphrase-based topic extraction for Urdu language and re-ranking of keyphrase has not applied before. The framework first extracts the positions of each keyword from the target document and ranks these keywords according to their positions. After ranking positional weights to each keyword in document, keyphrases of different sizes are generated by applying several syntactic rules. Furthermore, these generated keyphrases are ranked according to the scores of keyword and higher ranked keyphrases are extracted. For the extraction of more relevant keyphrases, a re-ranking of the keyphrases is introduced which extracts top-ranked keyphrases as potential topics of the target document where the one with the highest score is selected as the topic of the document. We have conducted experiments on two dataset of Urdu language which contains multiple documents from several different domains. Our framework produces better results and generates more meaningful topics for Urdu language as compared with existing techniques. Our system is capable to extract more accurate and meaningful topics and outperformed the existing approaches. Our methodology has some limitations like the efficiency and accuracy of the POS tagger etc. In future, the proposed approach will be enhanced by using graph based ranking of the keyphrases and improving the results of POS Tagger for Urdu language.

REFERENCES

[1] J. Peng, Y. Zhou, X. Sun, J. Su, and R. Ji, "Social media based topic modeling for smart campus: A deep topical correlation analysis method," *IEEE Access*, vol. 7, pp. 7555–7564, 2019, doi: 10.1109/ACCESS.2018.2890091.

[2] H. M. Wallach, "Topic modeling: Beyond bag-of-words," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, New York, NY, USA, 2006, pp. 977–984, doi: 10.1145/1143844.1143967.

[3] M. Reisenbichler and T. Reutterer, "Topic modeling in marketing: Recent advances and research opportunities," *J. Bus. Econ.*, vol. 89, no. 3, pp. 327–356, Apr. 2019, doi: 10.1007/s11573-018-0915-7.

[4] B. Chao and A. Sirmorya, "Automated movie genre classification with LDA-based topic modeling," *Int. J. Comput. Appl.*, vol. 145, no. 13, pp. 1–5, Jul. 2016, doi: 10.5120/ijca2016910822.

[5] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Beijing, China, vol. 1, 2015, pp. 1354–1364, doi: 10.3115/v1/P15-1131.

[6] T. Rana, Y.-N. Cheah, and S. Letchmunan, "Topic modeling in sentiment analysis: A systematic review," *J. ICT Res. Appl.*, vol. 10, no. 1, pp. 76–93, Jun. 2016, doi: 10.5614/itbj.ict.res.appl.2016.10.1.6.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, Sep. 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[8] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. Conf. Uncertainty Artif. Intell.*, San Francisco, CA, USA, 1999, pp. 289–296. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073796.2073829

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[10] K. Shakeel, G. R. Tahir, I. Tehseen, and M. Ali, "A framework of Urdu topic modeling using latent Dirichlet allocation (LDA)," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf. (CCWC)*, Las Vegas, NV, USA, Jan. 2018, pp. 117–123, doi: 10.1109/CCWC.2018.8301655.

[11] A. U. Rehman, Z. Rehman, J. Akram, W. Ali, M. A. Shah, and M. Salman, "Statistical topic modeling for Urdu text articles," in *Proc. 24th Int. Conf. Autom. Comput. (ICAC)*, Sep. 2018, pp. 1–6, doi: 10.23919/IConAC.2018.8748975.

[12] A. Ur Rehman, A. H. Khan, M. Aftab, Z. Rehman, and M. A. Shah, "Hierarchical topic modeling for Urdu text articles," in *Proc. 25th Int. Conf. Autom. Comput. (ICAC)*, Lancaster, U.K., Sep. 2019, pp. 1–6, doi: 10.23919/IConAC.2019.8895047.

[13] K. Khalid, H. Afzal, F. Moqaddas, N. Iltaf, A. M. Sheri, and R. Nawaz, "Extension of semantic based Urdu linguistic resources using natural language processing," in *Proc. IEEE 15th Int. Conf. Dependable, Autonomic Secure Comput., 15th Int. Conf Pervasive Intell. Comput., 3rd Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Nov. 2017, pp. 1322–1325, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.214.

[14] A. Daud, W. Khan, and D. Che, "Urdu language processing: A survey," *Artif. Intell. Rev.*, vol. 47, no. 3, pp. 279–311, Mar. 2017, doi: 10.1007/s10462-016-9482-x.

[15] K. Mehmood, D. Essam, and K. Shafi, "Sentiment analysis system for Roman Urdu," in *Intelligent Computing*. Cham, Switzerland: Springer, 2019, pp. 29–42, doi: 10.1007/978-3-030-01174-1_3.

[16] K. Ahmed, M. Ali, S. Khalid, and M. Kamran, "Framework for Urdu news headlines classification," *J. Appl. Comput. Sci. Math.*, vol. 10, no. 1, pp. 17–21, 2016, doi: 10.4316/JACSM.201601002.

[17] T. Zia, M. Akhter, and Q. Abbas, "Comparative study of feature selection approaches for urdu text categorization," *Malays. J. Comput. Sci.*, vol. 28, pp. 93–109, Jan. 2015.

[18] M. Usman, Z. Shafique, S. Ayub, and K. Malik, "Urdu text classification using majority voting," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, pp. 265–273, 2016, doi: 10.14569/IJACSA.2016.070836.

[19] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, A. Mehmood, and M. T. Sadiq, "Document-level text classification using single-layer multisize filters convolutional neural network," *IEEE Access*, vol. 8, pp. 42689–42707, 2020, doi: 10.1109/ACCESS.2020.2976744.

[20] M. P. Akhter, Z. Jiangbin, I. R. Naqvi, M. Abdelmajeed, and M. Fayyaz, "Exploring deep learning approaches for Urdu text classification in product manufacturing," *Enterprise Inf. Syst.*, early access, pp. 1–26, May 5, 2020, doi: 10.1080/17517575.2020.1755455.

[21] M. Huang, Y. Rao, Y. Liu, H. Xie, and F. L. Wang, "Siamese network-based supervised topic modeling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Brussels, Belgium, 2018, pp. 4652–4662, doi: 10.18653/v1/D18-1494.

[22] W. Li, J. Yin, and H. Chen, "Supervised topic modeling using hierarchical Dirichlet process-based inverse regression: Experiments on E-commerce applications," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1192–1205, Jun. 2018, doi: 10.1109/TKDE.2017.2786727.

[23] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, "Fast online EM for big topic modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 675–688, Mar. 2016, doi: 10.1109/TKDE.2015.2492565.

[24] X. Li, J. Ouyang, and X. Zhou, "Supervised topic models for multi-label classification," *Neurocomputing*, vol. 149, pp. 811–819, Feb. 2015, doi: 10.1016/j.neucom.2014.07.053.

[25] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Los Angeles, CA, USA, Jun. 2010, pp. 804–812. [Online]. Available: https://www.aclweb.org/anthology/N10-1122.

[26] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints," *Expert Syst. Appl.*, vol. 127, pp. 256–271, Aug. 2019, doi: 10.1016/j.eswa.2019.03.001.

[27] R. Venkatesaramani, D. Downey, B. Malin, and Y. Vorobeychik, "A semantic cover approach for topic modeling," in *Proc. 8th Joint Conf. Lexical Comput. Semantics (SEM)*, Minneapolis, Minnesota, 2019, pp. 92–102, doi: 10.18653/v1/S19-1011.

[28] X. Zhang and R. He, "Topic extraction of events on social media using reinforced knowledge," in *Proc. Int. Conf. Knowl. Sci. Eng. Manage.*, Changchun, China, vol. 11062, Aug. 2018, pp. 465–476, doi: 10.1007/978-3-319-99247-1_41.

[29] Z. Wang, K. Hahn, Y. Kim, S. Song, and J.-M. Seo, "A news-topic recommender system based on keywords extraction," *Multimedia Tools Appl.*, vol. 77, no. 4, pp. 4339–4353, Feb. 2018, doi: 10.1007/s11042-017-5513-0.

[30] M. Alhawarat and M. Hegazi, "Revisiting K-means and topic modeling, a comparison study to cluster Arabic documents," *IEEE Access*, vol. 6, pp. 42740–42749, Jul. 2018, doi: 10.1109/ACCESS.2018.2852648.

[31] A. Bougouin, F. Boudin, and B. Daille, "TopicRank: Graph-based topic ranking for keyphrase extraction," in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, Nagoya, Japan, Oct. 2013, pp. 543–551. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00917969.

[32] M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, and J. Han, "Automatic construction and ranking of topical keyphrases on collections of short documents," in *Proc. SIAM Int. Conf. Data Mining*, Apr. 2014, pp. 398–406.

[33] D. Parveen, H.-M. Ramsl, and M. Strube, "Topical coherence for graph-based extractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Lisbon, Portugal, 2015, pp. 1949–1954, doi: 10.18653/v1/D15-1226.

[34] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 2, 2018, pp. 667–672, doi: 10.18653/v1/N18-2105.

[35] M. R. Alfarra and A. Alfarra, "Graph-based technique for extracting keyphrases in a single-document (GTEK)," in *Proc. Int. Conf. Promising Electron. Technol. (ICPET)*, Deir El-Balah, Palestinian, Oct. 2018, pp. 92–97, doi: 10.1109/ICPET.2018.00023.

[36] X. Wan, J. Yang, and J. Xiao, "Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics*, Prague, Czech Republic, vol. 2, 2007, pp. 552–559. [Online]. Available: https://www.aclweb.org/anthology/papers/P/P07/P07-1070/.

[37] S. Danesh, T. Sumner, and J. H. Martin, "SGRank: Combining statistical and graphical methods to improve the state of the art in unsupervised keyphrase extraction," in *Proc. 4th Joint Conf. Lexical Comput. Semantics*, Denver, CO, USA, 2015, pp. 117–126, doi: 10.18653/v1/S15-1.

[38] C. B. Ali, R. Wang, and H. Haddad, "A two level keyphrase extraction approach," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, Cairo, Egypt, vol. 9042, 2015, pp. 390–401, doi: 10.1007/978-3-319-18117-2_29.

[39] C. Florescu and C. Caragea, "PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Vancouver, BC, Canada, vol. 1, 2017, pp. 1105–1115, doi: 10.18653/v1/P17-1102.

[40] I. Rasheed, H. Banka, and H. M. Khan, "A hybrid feature selection approach based on LSI for classification of Urdu text," in *Machine Learning Algorithms for Industrial Applications*, S. K. Das, S. P. Das, N. Dey, and A.-E. Hassanien, Eds. Cham, Switzerland: Springer, 2021, pp. 3–18.

AHMAD AMIN is currently pursuing the M.S. degree with the Computer Science and IT Department, The University of Lahore, Pakistan. His research interests are NLP and data mining. He is working under the supervision of Dr. Toqir A. Rana.

TOQIR A. RANA received the Ph.D. degree from Universiti Sains Malaysia. He is currently working as an Assistant Professor with the Computer Science and IT Department, The University of Lahore, Pakistan. His research interests are data mining, text mining, sentiment analysis, and NLP. He has published several articles in top journals and international conferences.

NATASH ALI MIAN received the M.C.S. degree from the University of Lahore, the M.S.(CS) degree from SZABIST, Islamabad, and the Ph.D. degree from NCBA&E, Lahore. He is currently working as an Assistant Professor with the School of Computer and Information Technology (SCIT), Beaconhouse National University. He specializes in software engineering with special interest in requirement engineering, self-adaptive systems, Internet of Things, cloud computing, formal methods, and reverse engineering and databases.

MUHAMMAD WASEEM IQBAL received the Ph.D. degree from The Superior College (University Campus) Lahore. He is currently working as the Head of Software Engineering Department, The Superior College (University Campus). He specializes in human computer interaction (HCI), with special interest in adaptive interfaces for mobile devices, usability evaluation of mobile devices for normal, and visually impaired people and user context ontological modeling.

ABBAS KHALID received the master's degree in computer science from the University of Central Punjab, Lahore, Pakistan, and the Ph.D. degree from Lancaster University, U.K. He is currently an Assistant Professor with the University of Lahore, Lahore. He has over 15 years of experience in academics and research. His research interests include communication systems, the Internet of Things, and robotics.

TAHIR ALYAS (Member, IEEE) received the master's degree in computer science and the Ph.D. degree from the School of Computer Science, NCBA&E, Lahore, Pakistan. He is currently working as the Head of the Department of Computer Science, Lahore Garrison University, Lahore. His research interests include cloud computing, fog computing, Hyper-convergence, the IoT, and intelligent age. He is also Oracle certified in Cloud Infrastructure Architect, Associate, Professional, and Oracle Autonomous Database Cloud 2019 Specialist.

MOHAMMAD TUBISHAT received the B.Sc. degree in computer science and the M.Sc. degree in computer and information sciences from Yarmouk University, Jordan, in 2002 and 2004, respectively, and the Ph.D. degree in computer science (artificial intelligence—natural language processing) from the University of Malaya, Malaysia, in 2019. He is currently working as a Lecturer with the Asia Pacific University of Technology and Innovation, Kuala Lumpur, Malaysia. His research interests include natural language processing, data mining, artificial intelligence, machine learning, optimization algorithms, data science, and sentiment analysis.