IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON SPECIAL SECTION ON ADVANCED SOFTWARE AND DATA ENGINEERING FOR SECURE SOCIETIES

Received November 7, 2018, accepted November 21, 2018, date of publication December 4, 2018,
date of current version January 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885011

An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution

WAHEED ANWAR[1], IMRAN SARWAR BAJWA[1], M. ABBAS CHOUDHARY[2], AND SHABANA RAMZAN[3]

[1]Department of Computer Science and Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[2]Dadabhoy Institute of Higher Education, Karachi 75400, Pakistan
[3]Department of Computer Science, Government Sadiq College Women University, Bahawalpur 63100, Pakistan

Corresponding author: Waheed Anwar (waheed@iub.edu.pk)

This work was supported by the U.S. Department of Commerce under Grant BS123456.

ABSTRACT In the recent years, text-based digital forensic has evolved into a major research domain that supports digital investigation. A piece of text can be a critical source of information that is written by somebody with respect to writing style, usage of typical vocabulary, and so on. In this paper, we present a unified approach for intelligent association analysis of text of how much a piece of text is related to a person with respect to his stylometric writing features. The latent Dirichlet allocation (LDA)-based approach emphasizes on instance-based and profile-based classification of an author's text. Here, LDA suitably handles the high dimensional and sparse data by allowing more expressive representation of text. The presented approach is an unsupervised computational methodology that can handle the heterogeneity of the dataset, diversity in writing styles of authors, and the inherent ambiguity of Urdu language text. A large corpus was collected for performance testing of the presented approach. The results of the experiments show the superiority of the proposed approach over the state-of-the-art representations and other algorithms used for authorship attribution. Manifold contributions of the presented paper are use of improved sqrt-cosine similarity with LDA topics to measure similarity in vectors of text documents for the forensic analysis purpose, construction of a large data set of 6000 documents of articles, and achievement of (92% f1-measure) results on articles without using any labels for authorship attribution task.

INDEX TERMS Authorship attribution, forensic analysis, computational linguistics, LDA, machine learning.

I. INTRODUCTION

Modern digital investigation relies on 'digital evidence' that plays an important role in investigations purpose and is considered an important piece of information. Typically, digital forensics aim extraction of information from the evidence data to answer the 5Ws (Why, When, Where, What, and Who). A digital forensic process assumes complete control of digital evidence and source of the evidence. A piece of text (such as customer reviews, reports, articles, etc.) can be a digital evidence and can be a source of serious information that can help in investigation. To find the writer or author of a piece of text is long-standing challenge as it needs a robust approach to investigate the feature based identification of author of a text [1]. This type of task involves association analysis that focuses on finding a set of patterns and features that describe the relationships among the binary attributes (variables) that are particularly used to characterize a set of objects. The stat-of-the-art perspectives of stylometry research is authorship analysis [2]–[7]. In the recent past, the domain of authorship analysis has embraced new dimensions of research typically with the emergence of machine learning techniques for text mining. One of the recent and emerging trends in authorship analysis is machine-based extraction of stylometric features from text of an author instead of manually engineering the stylometric features [8]–[10]. Authorship attribution is also attributed an important problem in domain of computational linguistics and information retrieval. The main focus of Authorship attribution is deciding the most appropriate author of a target document among a list of known author's [3]. From the machine learning aspect Authorship attribution can be perceived as one label multiclass text classification problem where the role of the classes are played by contestant authors [11].

The detailed literature in the domain of Authorship attribution for last two decades revealed that it is a field of great interest of the community and has been mainly applied on English language [4], [6], [12], [13]. Additionally, a few solitary efforts are done for application of Authorship attribution in other natural languages such as Greek [7], [14], Portuguese [15], [16], Dutch [17]–[19] and Arabic [6], [20]. However, during the literature review, it was found that there is no major contribution in the field of Authorship attribution of Urdu text except Urdu poetry [21]. To the best of our knowledge, neither a theoretical support nor a tool is available for Authorship attribution of Urdu newspaper columns that provides higher accuracy. There are more than 70 million native speakers of Urdu language in Pakistan, India, UAE, and in few other parts of the world. Additionally, Urdu has been medium of millions of books, manuscript, magazines, newspapers, etc. So, such Authorship attribution application for Urdu language is a real need of the time.

Latent Dirichlet Allocation (LDA) [22] has been found to be a flexible generative probabilistic unsupervised topic model typically used for the Authorship attribution for text documents [8], [9], [23], [24]. LDA has previously been used with similarity measuring technique such as Hellinger [9]. During the literature review, it was found that the results of the previously used similarity measuring techniques provide low accuracy and need improvement in topic matching process of LDA based author attribution. In this paper, we propose the use of improved sqrt-cosine distance metric with LDA topics to find similarity in vectors of text documents. In literature review, it was identified that the improved sqrt-cosine similarity (ICS) [25] has not been previously employed with LDA for Authorship attribution of the text documents. One of the objectives of the research presented in this paper was to investigate the behaviour of improved sqrt-cosine similarity with LDA in comparison with other similar previously used techniques for Authorship attribution.

The presented approach builds LDA models on n-grams texts instead of simple text, to keep personal stylistic attributes of the text writer and then improved sqrt-cosine similarity metric is used to find out similarity in LDA topical representation of Urdu text documents to carried out classification. Here, LDA's application on n-grams words not only keep various stylistic fingerprints to identify the writing style of a particular author but also can analyse a large dataset of Urdu newspaper articles and can identify the potential author for testing dataset. The presented approach emphasizes on author instance-based and profile-based classification of text. We used LDA which can handle high dimensional and sparse data, allowing more expressive representation of texts. LDA is also suitable considering the heterogeneity of the dataset, inherit ambiguity of Urdu language text and diversity in writing styles of authors. A large dataset was collected for performance testing of the presented approach. The results of experiments show superiority of the proposed approach over the state-of-the-art representations and other algorithms used for Authorship attribution. Manifold contributions of the presented work are use of improved sqrt-cosine similarity with LDA to measure similarity in vectors of text documents, construction of a large data set of 6000 documents of Urdu newspaper articles, and achievement of satisfactory results on Urdu news articles without using any labels for Authorship attribution task.

The rest of the paper is structured as follows: Section II discusses the outcomes of the detailed literature review carried out during the research and a course of related work; Section III describes the materials and methods of the presented research along the dataset collected and the LDA based used approach for Authorship attribution in Urdu newspaper articles; Section IV provides details of the experiments their results and discussions to show the performance testing and outcomes of the presented approach; Section V presents conclusion of the presented research. The paper ends with a conclusion section.

II. RELATED WORK

By using the stylometry and computational methods, data scientists and researchers are trying to bring revolutionary modifications for improved authorship analysis tasks like Authorship attribution, author verification and author profiling.

In Authorship attribution as a first main paradigm, we can apply univariate or multivariate measures that can reflect the style of a particular author. Studies on individual measures such as frequencies of specific word or letter occurrence [26], average word length, mean sentence length [27], lexical richness [11] and integrated syntactic graphs [28] has been done, however none of these individual measures prove satisfactory [29]. In multivariate approach basic intuition is to take documents as points in vector space, and by using some appropriate distance measure, assign the questioned document to the author whose documents are closest to the questioned document. One such approach is Delta [30], similarly other distance based similarity functions have been applied to diverse feature sets for Authorship attribution as well [4], [14], [21].

As a second main paradigm, we can apply machine learning techniques to find the most appropriate author of the given text. In this paradigm we can see individual author as one category, we then require to define a classification model by applying various features this model can now identify individual author category among possible authors. For the simplicity machine learning techniques are further separated into two sub categories, one is supervised and other is unsupervised. Supervised techniques are those in which author-class labels are involve for classification, while unsupervised techniques do classification without prior knowledge of author-class labels.

For Authorship attribution supervised techniques include neural networks [31], [32], linear discriminant analysis [33], decision trees [6] and support vector machines(SVMs) [13], [34], [35]. SVM outperformed other supervised techniques in

TABLE 1. Comparison of our approach with the related work.

Source	Features used	Classifier	Corpus
Michal Rosen-Zvi *et al.* [8]	author-topic model	Topic entropy	Collection of NIPS conference papers
Abbasi and Chen[6]	Syntactic, structural and lexical features	SVM & decision trees	English and Arabic web posts
Koppel *et al.* [34]	Tf-idf over characters and words	SVM	English web posts
Stamatatos [35]	Character level n-grams	SVM	English and Arabic journalism
Zaho *et al.* [39]	Latent dirichlet allocation topics	Topic assignment	English social media
Seroussi[9]	Latent dirichlet allocation topics	Hellinger distance	Web posts and English e-mails
Savoy [24]	Latent dirichlet allocation topics	SVM	English and Italian journalism
Caliskan-Islam *et al.*[10]	Syntactic and Lexical features	Random forest	Source and compiled code
Proposed approach	**Words n-grams and latent dirichlet allocation topics**	**Improved sqrt-cosine similarity**	**Urdu journalism**

TABLE 2. Distribution of 6000 Urdu documents by author name, number of articles words and average words per document.

	Name	Number	Words	Avg. words
1	Abdul Qadir Hassan	400	418265	1046
2	Aftab Ahmed Khanzada	400	484256	1211
3	Asad Ullah Ghalib	400	474673	1187
4	Dr M. Ajmal Niazi	400	471024	1178
5	Dr Tauseef Ahmad Khan	400	526201	1316
6	Haroon Ur Rashid	400	534802	1337
7	Irshad Ahmad Arif	400	571798	1430
8	Irshad Ansari	400	158309	396
9	Javed Chaudhary	400	676141	1690
10	Karnal Ikram Ullah	400	345903	865
11	Khursheed Nadeem	400	511401	1279
12	Nawaz Raza	400	268674	672
13	Nazeer Naji	400	590991	1478
14	Qayyum Nizami	400	501933	1255
15	Zahida Hina	400	603032	1508

head-to-head comparisons such as decision trees and neural networks.

Unsupervised classification techniques include cluster analysis [36], principal component analysis (PCA) [37], [38] and LDA [8], [9], [23], [24]. The pioneer systematic study of Authorship attribution by using extended version of LDA was introduce by Michal Rosen-Zvi *et al.* [8]. LDA's ability to capture all hidden topics from large numbers of features in a reduced dimensionality makes it appealing for text analysis problems.

An overview of the closely related work that used various features sets and classifiers for author attribution is shown in TABLE 1.

III. MATERIALS AND METHODS

In this section major steps of our proposed framework for Authorship attribution are discussed, with the intention to explain the necessary information related to our corpus, datasets, algorithms, models and their specific parameter settings and experiments, so that results can be reproducible. The materials used is corpus in Urdu language TABLE 2 datasets TABLE 3 and inferred topics, the methods were data collection Section III-A, pre-processing Section III-B, various features extraction and selection Sections III-C, document term matrix preparation Section III-D, topics extraction using Latent Dirichlet Allocation Section III-E, proposed LDA + improved sqrt-cosine methodology for classification Section III-F.

TABLE 3. Datasets used in the experiments.

Description	Training Documents	Testing Documents
instance-based	4500	1500
instance-based with n-grams	4500	1500
profile-based	15	1500
profile-based with n-grams	15	1500

A. CORPUS

In the Authorship attribution domain, there are two issues regarding corpus, one the number of publicly available test corpora are quite limited two, the available corpora are of comparatively small in term of number of texts documents or in term of number of authors. Thus, producing adequately precise comparisons between reported performances is problematic.

To the best of our knowledge there is no benchmark corpus for authorship in Urdu language until now. We decided to build new corpus in Urdu language, we used Urdu articles from news domain for this corpus. We wrote webpage scraping scripts in PHP programming language for each newspaper as the webpage structure of each newspaper was diverse. The process of data extraction was in two steps in first step, all the URLs of specific author were extracted by using crawler and in second step webpage scraper used these URLs to extract all available article contents FIGURE 1. We initially collected over 21,938 articles from main stream Urdu newspapers of Pakistan namely Express [40], Nawa-e-waqat [41] and Dunya [42].

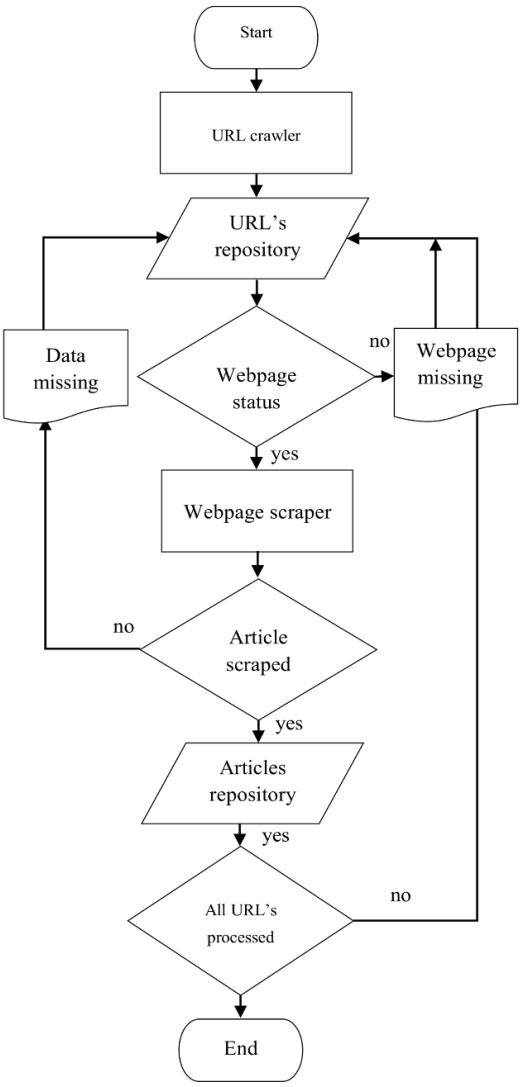


FIGURE 1. Data collection process from websites using web crawler and web scraper.

No additional contents were added or deleted from the articles except html tags were removed. There was no limit imposed on maximum number of words in downloaded article. Since more information about word structuring and writing style provides, better trained models, leading towards better prediction and accurate results. However minimum length of an article to be included in corpus was set to be 85 words, since it is harder to extract stylistic and content-based features from short article. In TABLE 2 we selected 6,000 articles of 15 authors with 400 articles for each author, for the sake of balance corpus, naming it UrduCorpus. We shall publicly share this corpus and its various dataset used here.

UrduCorpus contains 7,137,403 tokens in total, at the document level, the mean length was 1189 tokens. The longest document was written by Aftab Ahmed Khanzada (2,223 tokens) and the shortest by Irshad Ansari (86 tokens). When considering the mean length per author, Irshad Ansari wrote the shortest documents (396 tokens per document) while Javed Chaudhary is the author, of the longest documents (1,690 tokens per document).

1) DATASETS

In datasets preparation from UrduCorpus, we used two representations of author specific documents.

a: INSTANCE-BASED

In instance-based representation all documents were treated individually.

b: PROFILE-BASED

All the author specific documents were concatenated into single file, now this single document shall represent individual author, in this way we have only fifteen long concatenated documents in total each one representing unique author.

We prepared four datasets from UrduCorpus as shown in TABLE 3. among these datasets two were instance-based with and without n-grams and two were profile-based with and without n-grams. We used randomly 75% data for training and 25% data for testing with respect to each author.

Two profile-based datasets of UrduCorpus have only fifteen (15) lengthy documents for training on the other hand each dataset have equal test documents for model evaluation.

The proposed framework for Authorship attribution FIGURE 2. Using topic modeling with LDA with improved sqrt-cosine similarity for classification.

B. PRE-PROCESSING

It is observed from literature review that it is not needed for vigorous pre-processing in Authorship attribution. As writer's grammatical mistakes, their preferences of letter abbreviation, letter capitalization, word prefixes and suffixes all are essential part of one's writing style. In this case, it is not feasible to correct grammatical mistakes or stem words, such actions may reduce the number of features specific to writer. We used Natural Language Tool Kit (NLTK) [43] for tokenizing at word-level after ignoring all whitespaces. We trimmed the phrases that appeared in fewer than 10 documents and more than 90% of the documents in the UrduCorpus.

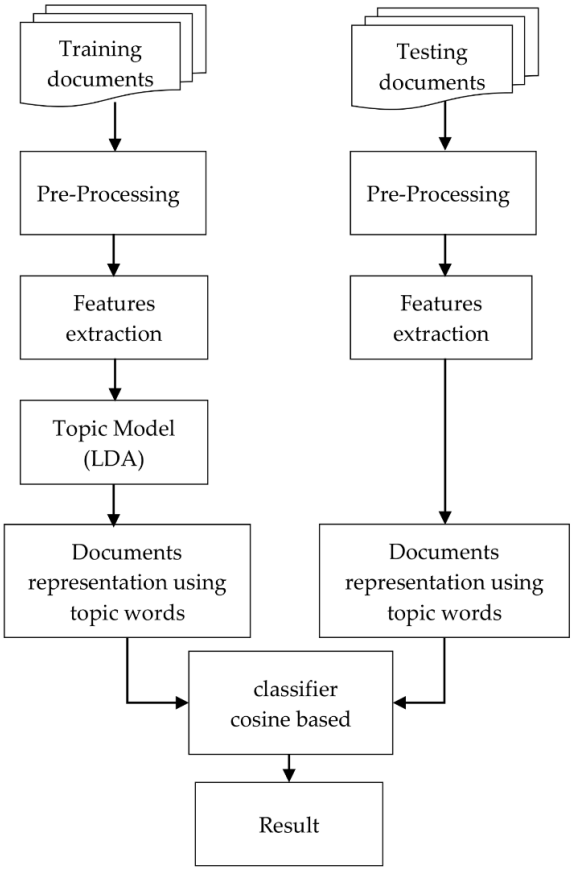


FIGURE 2. Proposed framework for authorship attribution.

C. FEATURE EXTRATION

To extract numerical information from raw text documents is normally termed as Feature extraction process. Among extracted features only those features are selected that best fit the training model. After this process, if the features set dimensionality is huge and difficult for computation then it requires dimensionality reduction algorithms for appropriate performance. The following feature extracting techniques were used for proposed model.

1) BAG OF WORDS

In natural language processing bag-of-words is a classic model. In this model text is considered as a set of words each one having a frequency of occurrence in the corpus, however their contextual information is lost. In other words, it is order less document features representation in the form of frequencies that occurs in the document to form a dictionary, this dictionary may consist of character, character n-grams, words, words n-grams, or some other features extracted from text. We can produce distinct feature vectors based on information captured from the texts. This could be simply raw frequency of each word or term frequency and inverse document frequency(tf-idf). In this article we used raw frequencies of words at corpus level.

2) N-GRAMS

In any text document n-grams are all groupings of adjacent characters or words of length n. These n-grams features are language independent we can capture them from any language. From statistical prospective they can capture the language structure of a writer, like what character or word was expected to follow the given one. The choice of n is very important in n-grams, if it produces short n-grams we may fail to capture important differences. On the other hand, if it produces long n-grams we may only stick to particular cases. Optimum length really depends on the application, a good rule of thumb in word level n-grams is to use n-grams where $n \in \{1, \ldots, 5\}$, this will significantly increases the length of the feature vectors almost five time as compare to normal vectors of the documents.

To overcome bag of word model limitation of contextual information lost, with n-grams we can capture more semantically meaningful information from text. Lexical n-grams are becoming popular as they are shown to be more effective than character n-grams and syntactic n-grams when all the possible n-grams are used as features [44] Moreover, it has been shown to be effective in identifying the gender of tweeters [45]. For ease of understanding we used underscores (_) to replace spaces in word n-grams and represent them as a single word in the vocabulary and subsequently in the bag of word model. The subsequent example shows a simple sentence and its complete lists of unigrams, bigrams, trigrams, fourgrams and fivegrams words generated from it. Note that Urdu is written from right to left, so read following sentence from right to left and n-grams from left to right.

Text: (یہ اللہ تعالی کا احسان ہے)
Unigrams:
("یہ", " اللہ", " تعالی", " کا", "احسان", "ہے")
Bigrams:
("یہ_اللہ", " اللہ_تعالی", " تعالی _کا", "کا_احسان", "احسان_ہے")
Trigrams:
("یہ_اللہ_تعالی", "اللہ_تعالی_کا", " تعالی _کا_احسان", "کا_احسان_ہے")
Fourgrams:
("یہ_اللہ_تعالی_کا", "اللہ_تعالی_کا_احسان", "تعالی _کا_احسان_ہے")
Fivegrams:
("یہ_اللہ_تعالی_کا_احسان", "اللہ_تعالی_کا_احسان_ہے")

For word-level n-grams feature vector length varies as choice of n varies, it can grow rapidly almost n-times with n-grams. We find 7,137,403 words in training documents and for the sake of defining the style of particular author we applied n-grams where n=5 now the size of word types were 35,687,005 words in total, we find 12,411,430 distinct words among these words first we want to ignore words with small frequency occurrence as they represent a large proportion of the vocabulary. There were 11,048,391 hapax legomenon (words occurring once) and 722,907 dis legomenon (words occurring twice). If we use all distinct words for our vocabulary it can increase overall corpus dimensionality which is difficult for computation. As a feature selection we have applied a scheme of considering only those terms having frequency

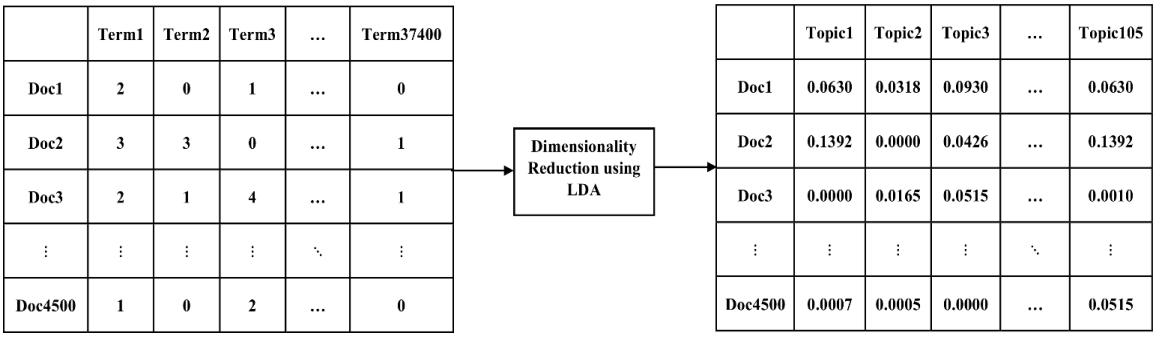


FIGURE 3. Conversion of document term matrix to document topic matrix.

occurrence of 30 or more ($tf \geq 30$), it reduced vocabulary size to 38121 terms. We also add the second constraint that each selected word should not appear in every document Thus, we want to ignore stop words appearing in almost every document. we ignore all words occurring in 70 percent or more documents. Taking into account this second constrain we ignore 686 most frequent words having frequency range from 901 to 4,448. Finally, we obtain a vocabulary of size 37,444 terms in instance-based n-grams dataset, similar feature selection schemes were applied on simple instance and profile-based datasets we obtained vocabulary of size 14,659 terms and on profile-based n-grams dataset applying slightly different feature selection scheme we capture 75,125 terms for vocabulary.

D. PREPARING DOCUMENT TERM MATRIX

Text documents are generally represented as a vector where in a document each attribute represents particular term frequency occurrence. This vector form representation can be used to find the similarity between the two corresponding documents. It is recommended to convert text corpus into a matrix representation before running any mathematical model on it. We prepared document term matrix

FIGURE 3 from training dataset based on selected features which were saved in the form of vocabulary by using gensim dictionary class. LDA model looks for repeating term patterns in the entire document term matrix.

E. TOPIC MODELLING USING LDA

We can use topic models for the purpose of information retrieval and feature selection from unstructured text. A Topic modeling algorithm, say Latent Dirichlet Allocation [22] is useful for organizing large volume of textual data into overlapping clustering of documents [22], [46] which differ from other text mining approaches, which are rule-based and use dictionary or regular expressions-based keyword searching. LDA is a flexible generative probabilistic topic model for collection of discrete data, that express the documents as collection of a mixture of topics with different probabilities for these topics in documents, and each topic is expressed as list of words with probabilities for them to belong to that topic. However, a selected document may consist of only on single topic or multiple topics with different proportion. For example, if we have three documents d1, d2 and d3 in the whole corpus and we want to generate three topics t1, t2 and t3 the document d1 may have topic t1 intensively, some proportion related to topic t2 and little bit of topic t3. The document d2 have equal mixture of topic t1 and t2 and document d3 may have only topic t3.

Though, LDA does not directly provide author of the document; however, it can still be used to capture valuable information about the writer of the document. The k topics, output of LDA, are usually much smaller than the size of the vocabulary V.

We have used LDA to reduce dimension of document term matrix FIGURE 3 to new matrix, we named new matrix as document topic matrix, as each cell represents specific topic weightage in that document and each matrix row now have topical representation of whole document in normalized form so, these topical representations formed dimensionality reduction of the document term matrix from 4500×37400 to document topic matrix 4500×105 which is almost 99% less of the vocabulary of the corpus, which is extremely helpful in features selection and classification of documents.

1) HYPER PARAMETERS & PARAMETERS OF LDA

LDA has corpus level parameters named hyper parameters α and β sampled only once, these parameters are from dirichlet distribution. First parameter α controls per document topics dispersion and β accountable per topic words dispersion, the high value of α mean each document possibly have mixture of almost every topic not a particular topic while low value of α mean document is represented by some of topics, similarly high β value mean each topic is possibly have mixture of most of the words not just specific words while low value of β mean a topic may represent a blend of just some of words. In a nutshell high α will produce documents more identical to each other and high β will produce topics more identical to each other. Literature analysis shows that final research goal plays important role in LDA values as well as procedures of optimization for LDA hyper parameters α and β, as these parameters affect sparsity of the document

topics and topic word distributions. we have preferred to fix hyper parameters throughout all experiments both default to a symmetric 1.0/number of topics prior. Whereas number of topics k is user defined we need to figure out the number of topics based on the data. Thus each document di, for $i = 1, \ldots,$ n, is generated based on a distribution over the k topics, where k defines the maximum number of topics. This value is fixed and defined a priori, a lower value for the number of topics may results in border topics like education, sports and fashion, and a larger value for number of topics may results in more focused topics like science, football and hairstyle. A large k value for topics means that the algorithm requires lengthier passes to estimate the word distribution for all the topics, a good rule of thumb is to choose a value that make sense for particular case, in the present context, we may consider k =15 at least assuming that each k value corresponds to the individual author writing style and thus choosing $k = 15$ was a sensible choice, larger k value than 15 was required to imitate the versatility of writing style of a particular author as two or more topics distributions were more helpful in this regard. however lower k value than 15 does not make any sense. We used k between 15 and 105 with the interval of 10.

F. LDA + IMPROVED SQRT-COSINE SIMILARITY

This method is our main contribution, as it achieves state-of-the-art performance in Authorship attribution with many candidate authors. The main idea of our approach is to use LDA model in such a way that it provides us dimensionality reduction along with maintaining the author specific writer style, then use improved sqrt-cosine similarity in LDA model topic space, to determine the most likely author of the test document. We used n-grams to capture the author writing style. Documents were represented as bag-of-words, so each document from both training and test sets converted into sparse vector and were mapped into LDA topic space to generate a vector representation for each one, which can be represented as u_i and v_i as outcomes respectively.

In text similarity measures cosine similarity is one of the most popular one. It is a distance metric from computational linguistics to measure similarity between documents vectors. In order to find cosine similarity between two documents u and v first we need to normalize them to one in L_2 norm.

$$\sum_{i=1}^{k} u_i^2 = 1 \tag{1}$$

Now cosine similarity between these two normalized vectors u and v will be the dot product of them.

$$Cos(u, v) = \frac{\sum_{i=1}^{k} u_i v_i}{\sqrt{\left(\sum_{i=1}^{k} u_i^2\right)}\sqrt{\left(\sum_{i=1}^{k} v_i^2\right)}} \tag{2}$$

Zhu *et al.* [47] proposed a new cosine similarity measure sqrt-cosine similarity based on Zhu *et al.* [47] proposed a new cosine similarity measure sqrt-cosine similarity based on L_1 norm as the L_2 norm, is not a required metric for high-dimensional data mining application [48]. Based on these experiments, In Equation (3), each document is normalized to 1 in L_1 norm: $\sum_{i=1}^{k} u_{i=1}$.

$$SqrtCos\,(u, v) = \frac{\sum_{i=1}^{k} \sqrt{u_i v_i}}{\left(\sum_{i=1}^{k} u_i\right)\left(\sum_{i=1}^{k} v_i\right)} \tag{3}$$

Sqrt-cosine similarity in some cases, conflict with the definition of similarity measurement [25].

To use the benefits of sqrt-cosine similarity another similarity measure based on same philosophy, improved sqrt-cosine similarity (ISC) [25] was proposed. In this measure instead of using L_1 norm, square root of L_1 norm was used Equation (4). It has a non-negative value and bounded between [0, 1] which is better assessed with probability-based approaches.

$$ISC(u, v) = \frac{\sum_{i=1}^{k} \sqrt{u_i v_i}}{\sqrt{\left(\sum_{i=1}^{k} u_i\right)}\sqrt{\left(\sum_{i=1}^{k} v_i\right)}} \tag{4}$$

where u_i and v_i are vectors of n-dimensions over the document set **u** and **v** where $i = 1, 2, 3, \ldots$,k.

Cosine similarity is considered as the one of the best in similarity measurement. Improved sqrt-cosine similarity is very similar to cosine similarity in implementation complexity such as in Gensim [49]. We also used different evaluation metrics in order to validate and compare our results.

IV. RESULTS AND DISCUSSIONS

In our experiments, we validated the proposed Authorship attribution scheme by preforming tests on four datasets of UrduCorpus. In order to build the low-dimensionality topical representation, the LDA model receives tokenized text documents with n-grams of the training set without any label (without the author to which they belong) as input data type and for evaluation the unlabeled text documents from the testing set. The experiments comprised in testing a cosine base classifier with the output of LDA k-topics in the corpus, these topics form a lower-dimensional representation of the corresponding training set based on vocabulary and then evaluating the classifier with the testing set using the same lower-dimensional representation. The overall Authorship attribution accuracy rate (AR) is computed by Equation (5) as below:

$$AR = \frac{number\ of\ correctly\ identified\ articles}{Total\ number\ of\ test\ set\ articles} \times 100 \tag{5}$$

A. EXPERIMENTAL SETUP

All the experiments were performed to test the performance and accuracy of the proposed approach using Intel i7 @ 2.8GHz, operating on windows 10 pro 64-bit with 6 GB memory. Python 3 (Python software Foundation, Wilmington, DE, USA), NLTK [43] and LDA implementation in Gensim [49] library were used for the development of system. LDA implementation in Gensim allows both estimation of

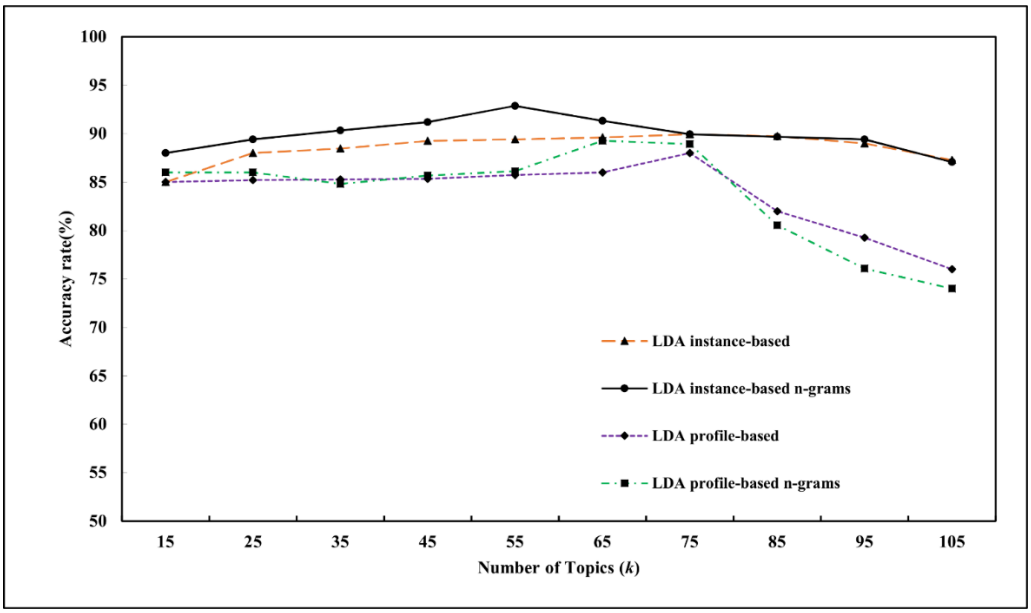


FIGURE 4. Evaluation of LDA authorship attribution using improved sqrt-cosine similarity on four datasets.

topics distribution on training data and inference of these topics on test data, for its parameters setting Section-III-D1.

We used UrduCorpus dataset TABLE 3 that belongs to the news domain. Note that, change of newspaper may affects writing style of an author similarly over the passage of time individual writing style may also change. The nature of articles (their topics) also influences the choice of words, however every individual has its own vocabulary it may like to use specific words unintentionally which can be used for its writing style identification.

To evaluate and compare LDA for Authorship attribution, we used the 6000 Urdu documents written by 15 well-known Urdu newspaper columnists. We used various performance metrics (Precision, Recall and f1-measure) along with accuracy to demonstrate the quality of auto decision making of cosine-based classifier on UrduCorpus.

B. RESULTS AND DISCUSSION

In order to substantiate the results we evaluated LDA-based Authorship attribution approach on instance and profile-based datasets with and without n-grams, we carried out a series of experiments on each dataset with several filters on term frequency as well as frequent words removal to generate vocabulary with most appropriate features and different number of LDA topics (15, 25, 35, . . . , 105). We presented each experiment with best performance parameter setting as shown in TABLE 4.

Our results clearly shows that LDA instance-based n-grams approach outperform LDA profile-based approach significantly, although we were hoping vice versa as mentioned in literature [9]. In profile-based approach when documents were concatenated into single file to form author profile, some important Authorship features lose their prominence

TABLE 4. Unsupervised classification of documents based on LDA topics with Improved sqrt-cosine similarity on four datasets of UrduCorpus.

Method & Dataset	Parameters	Accuracy rate (%)
LDA instance-based	Vocabulary 14659, k = 75	89.93
LDA instance-based with n-grams	Vocabulary 37444, k = 55	**92.89**
LDA profile-based	Vocabulary 14659, k = 75	88.00
LDA profile-based with n-grams	Vocabulary 75125, k = 65	89.26

in the profile, these features have significant discriminating power that sharply contrast documents between the authors. Secondly although we have used balanced corpus in term of number of documents but the average document length per author vary so, when concatenating documents into author profile some profiles have less number of words in total as compared to other resulting in unbalanced features extraction, whereas in instance-based approach some documents of an author were long enough to become strong candidate of attributed document. Thirdly in instance-based approach different features can be combined easily whereas in profile-based approach it is difficult to do so.

FIGURE 4 depicts the result of multiple experiments that compare the unsupervised classification of documents based on LDA topics with Improved sqrt-cosine similarity on four datasets of Section III.

We have reported the accuracy percentage yielded by the LDA + improved sqrt-cosine similarity approach, in LDA model setting the number of topics k between (15, 25, 35, 45,. . . , 105) with various vocabulary sizes Section III-C2. Our result shows that varying the number of topics in LDA model

TABLE 5. Unsupervised classification of author documents on instance-based n-grams dataset.

Authors	Precision	Recall	F1-measure	Support
Abdul Qadir Hassan	0.83	0.91	0.87	100
Aftab Ahmed Khanzada	0.93	0.90	0.91	100
Asad Ullah Ghalib	0.95	0.98	0.97	100
Dr. Muhammad Ajmal Niazi	1.00	0.98	0.99	100
Dr. Tauseef Ahmad Khan	0.97	0.99	0.98	100
Haroon Ur Rashid	0.95	0.96	0.96	100
Irshad Ahmad Arif	0.84	0.85	0.85	100
Irshad Ansari	1.00	0.98	0.99	100
Javed Chaudhary	0.98	0.99	0.99	100
Karnal R Ikram Ullah	0.96	0.91	0.93	100
Khursheed Nadeem	0.95	0.98	0.97	100
Nawaz Raza	0.98	0.86	0.91	100
Nazeer Naji	0.79	0.81	0.80	100
Qayyum Nizami	1.00	0.94	0.97	100
Zahida Hina	0.84	0.89	0.86	100
Average / total	**0.931**	**0.929**	**0.930**	**1500**

is critical, it has a huge impact on performance. Usually, accuracy increases with the number of topics in a certain range and then begins to decrease. A clear and precise prescription for this parameter is not possible, even in same dataset with different vocabulary sizes.

In order to evaluate the proposed LDA-based approach on four datasets. We used the same number of topics with identical vocabulary size initially, however the results were not satisfactory for couple of datasets, as with combination of n-grams document size increases in term of tokens and length in dataset, thus in these datasets we cannot use the same vocabulary size for each LDA model. We tuned LDA models with different vocabulary sizes keeping same k topics. We have reported the best performance of each dataset with different vocabulary sizes but same number of topics between 15 and 105, in the current context, we may assume that each topic at least matches to the writing style of an author thus fixing k =15 is a reasonable choice. However, the value of k could be larger than 15 representing the fact that each author may require two or more topical representation to well describe the style of a given author.

When applying the LDA model on instance-based n-grams dataset with vocabulary of 37444 terms and k=55 topics, we achieved an accuracy 92.8% with improved sqrt-cosine similarity. Hence evaluations reported in this graph indicate that the LDA-based Authorship attribution model performs significantly better on instance-based n-grams dataset than other datasets almost on each k topics selection. Note that to further elaborate the results in the following we used proposed model with instance-based n-grams dataset. FIGURE 5 shows confusion matrix obtained with proposed methodology on 1500 test documents.

This confusion matrix can be used for various performance measures which can evaluate our results in different ways. As we can see, there is a clear diagonal heatmap which represents the accuracy with respect to author, however there were some documents which were misclassified. Four out

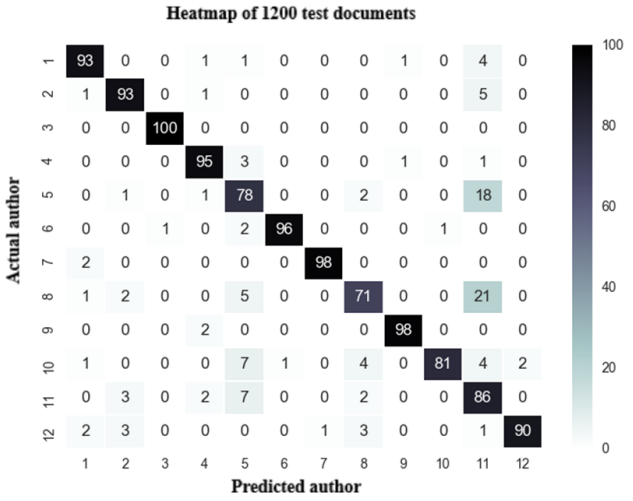


FIGURE 5. Confusion matrix for test documents using instance-based n-grams approach.

of fifteen authors have at least five misclassified documents towards single author for example eleven (11) documents for actual author number thirteen (13) were misclassified towards predicated author number one (1) which shows some resembles of their writing styles. One notable result was that authors with maximum accuracy also did not have any misclassified document in their favor, which shows their unique writing style.

1) INTERPRETATION OF MISCLASSIFIED ARTICLES

There can be number of reasons for misclassification of articles. Firstly, we found that few authors have writing style such that, in their articles first they gave quoted paragraphs of other authors and then discuss their point of view on that topic. In this way they intermingle their writing style with other authors. Secondly, authors wrote on various domains

like politics, religion, sports and entertainments as the corpus was not domain specific. It was possible some authors write more on specific domain hence; our proposed scheme may model that author somehow more towards that domain in consequence document of any other author of specific domain may misclassified. Thirdly, short size of the testing article may be the cause of misclassification.

In TABLE 5 we reported individual class results in terms of precision, recall, f1-measure and accuracy rate (percentage of correct answers) obtained on instance-based n-grams dataset by applying proposed scheme for Authorship attribution.

This experiment shows our approach models the authors more accurately on n-grams instance-based dataset. We achieved 92.89% accuracy rate on this dataset, other performance measures were also satisfactory, as precision measure was fluctuating from 79% to 100% and recall measure was between 81% to 99% on individual basis of this dataset. As there is a tradeoff between precision and recall, we attained 93.1% precision and 92.9% recall on 1500 test documents of instance-based n-grams dataset.

V. CONCLUSION

In this paper we designated the Authorship attribution problem in articles for efficient forensic analysis. As a new Authorship attribution scheme, we proposed an approach using Latent Dirichlet Allocation (LDA) paradigm in conjunction with n-grams to produce reduced dimension topical representation of UrduCorpus. We explained how the topical representations of LDA could be used with improved sqrt-cosine distance metric for classification of test documents. Our approach yields satisfactory performance. The best result in terms of accuracy and f1-measure were achieved with n-grams introduction in the model which captures more stylistic features of an author. The lessons learned were that each language required different configuration at each stage, appropriate selection of the dimensionality of the representation is crucial for Authorship attribution, and it is possible to significantly improve the accuracy results by fine tuning the size of vocabulary and k topics in LDA.

One possible improvement to the study would be implementation of supervised learning model to get good accuracy. This would increase the effort of annotating the corpus. Secondly, we could train the model developed in the study, on a larger set of columnists. One could aim to design and deploy an automated websites scraper incorporated with the proposed LDA model to collect other such online articles and create a comprehensive database of all such columnists. By doing so it could probably help Authorship attribution on a larger scale.

ACKNOWLEDGMENTS

The authors would like to thank Dr. M. Omer for his technical guidelines.

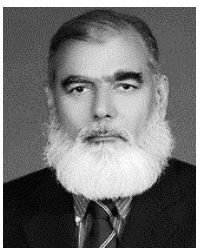
REFERENCES

[1] D. I. Holmes, "The evolution of stylometry in humanities scholarship," *Literary Linguistic Comput.*, vol. 13, no. 3, pp. 111–117, 1998.
[2] D. I. Holmes, "Authorship attribution," *Comput. Hum.*, vol. 28, no. 2, pp. 87–106, 1994.
[3] P. Juola, "Authorship attribution," *Found. Trends Inf. Retr.*, vol. 1, no. 3, pp. 233–334, 2008.
[4] C. E. Chaski, "Empirical evaluations of language-based author identification techniques," *Forensic Linguistic*, vol. 8, no. 1, pp. 1–65, 2001.
[5] M. Koppel, J. Schler, and S. Argamon, "Authorship attribution in the wild," *Lang. Resour. Eval.*, vol. 45, no. 1, pp. 83–94, 2011.
[6] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group Web forum messages," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 67–75, Sep. 2005.
[7] E. Stamatatos, N. Fakotakis, and G. Kokkinakis, "Text genre detection using common word frequencies," in *Proc. 18th Conf. Comput. Linguistic*, vol. 2, 2000, pp. 808–814.
[8] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *Proc. 20th Conf. Uncertain. Artif. Intell.*, 2004, pp. 487–494.
[9] Y. Seroussi, I. Zukerman, and F. Bohnert, "Authorship attribution with latent Dirichlet allocation," in *Proc. 15th Conf. Comput. Nat. Lang. Learn.*, Jun. 2011, pp. 181–189.
[10] A. Caliskan-Islam, "Stylometric fingerprints and privacy behavior in textual data," Ph.D. dissertation, Dept. Comput. Sci., Drexel Univ., Philadelphia, PA, USA, May 2015, p. 170.
[11] F. Sebastiani, "Machine learning in automated text categorization," *ACM Comput. Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
[12] D. Holmes, M. Robertson, and R. Paez, "Stephen crane and the *New-York Tribune*: A case study in traditional and non-traditional authorship attribution," *Comput. Hum.*, vol. 35, no. 3, pp. 315–331, 2001.
[13] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 9–26, 2009.
[14] V. Kešelj, F. Peng, N. Cercone, and C. Thomas, "N-Gram-based author profiles for authorship attribution," in *Proc. Pacific Assoc. Comput. Linguistic*, 2003, pp. 255–264.
[15] D. Pavelec, L. Oliveira, E. Justino, and L. Batista, "Using conjunctions and adverbs for author verification," *J. Univers. Comput. Sci.*, vol. 14, no. 18, pp. 2967–2981, 2008.
[16] R. S. Silva, G. Laboreiro, L. Sarmento, T. Grant, E. Oliveira, and B. Maia, "'twazn me!!!;('automatic authorship analysis of micro-blogging messages," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Berlin, Germany: Springer, Jun. 2011, pp. 161–168.
[17] P. Juola and R. H. Baayen, "A controlled-corpus experiment in authorship identification by cross-entropy," *Literary Linguistic Comput.*, vol. 20, no. 1, pp. 59–67, 2005.
[18] J. Hoorn, S. Frank, W. Kowalczyk, and F. van der Ham, "Neural network identification of poets using letter sequences," *Literary Linguistic Comput.*, vol. 14, no. 3, pp. 311–338, 1999.
[19] P. Maitra, S. Ghosh, and D. Das, "Authorship verification—An approach based on random forest notebook for PAN at CLEF 2015," in *Proc. Workshop Notes CLEF Conf.*, 2015, pp. 1–9.
[20] A. S. Altheneyan and M. El B. Menai, "Naïve Bayes classifiers for authorship attribution of Arabic texts," *J. King Saud Univ., Comput. Inf. Sci.*, vol. 26, no. 4, pp. 473–484, 2014.
[21] A. A. Raza, A. Athar, and S. Nadeem, "N-gram based authorship attribution in urdu poetry," in *Proc. Conf. Lang., Technol.*, 2009, pp. 88–93.
[22] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[23] R. Arun, R. Saradha, and V. Suresh, "Stopwords and stylometry: A latent Dirichlet allocation approach," in *Proc. NIPS Workshop*, 2009, pp. 1–4.
[24] J. Savoy, "Authorship attribution based on a probabilistic topic model," *Inf. Process. Manage.*, vol. 49, no. 1, pp. 341–354, 2013.
[25] S. Sohangir and D. Wang, "Document understanding using improved Sqrt-cosine similarity," in *Proc. IEEE 11th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2017, pp. 278–279.
[26] T. C. Mendenhall, "The characteristic curves of composition," *Science*, vol. 11, no. 214, pp. 237–249, 1887.
[27] G. Yule, "The statistical study of literary vocabulary," *Mod. Lang. Rev.*, vol. 39, no. 3, pp. 291–293, 1944.

[28] H. Gómez-Adorno, G. Sidorov, D. Pinto, D. Vilariño, and A. Gelbukh, "Automatic authorship detection using textual patterns extracted from integrated syntactic graphs," *Sensors*, vol. 16, no. 9, pp. 1–19, 2016.
[29] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," *Literary Linguistic Comput.*, vol. 22, no. 3, pp. 251–270, 2007.
[30] J. Burrows, "'Delta': A measure of stylistic difference and a guide to likely authorship," *Literary Linguistic Comput.*, vol. 17, no. 3, pp. 267–287, 2002.
[31] F. J. Tweedie, S. Singh, and D. I. Holmes, "Neural network applications in stylometry: The *Federalist Papers*," *Comput. Hum.*, vol. 30, no. 1, pp. 1–10, 1996.
[32] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, 2006.
[33] C. E. Chaski, "Who's at the keyboard: Authorship attribution in digital evidence investigations," *Int. J. Digit. Evidence*, vol. 4, no. 1, pp. 1–13, 2005.
[34] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Commun. ACM*, vol. 5, no. 2, pp. 119–123, 2006.
[35] E. Stamatatos, "Author identification: Using text sampling to handle the class imbalance problem," *Inf. Process. Manag.*, vol. 44, no. 2, pp. 790–799, 2008.
[36] D. Holmes, "A stylometric analysis of Mormon scripture and related texts," *J. Roy. Stat. Soc. A, Statist. Soc.*, vol. 155, no. 1, pp. 91–120, 1992.
[37] J. F. Burrows, "Word-patterns and story-shapes: The statistical analysis of narrative style," *Literary Linguistic Comput.*, vol. 2, no. 2, pp. 61–70, 1987.
[38] A. Jamak, A. Savatić, and M. Can, "Principal component analysis for authorship attribution," *Bus. Syst. Res.*, vol. 3, no. 2, pp. 49–56, 2012.
[39] W. X. Zhao, J. Jiang, J. Weng, J. He, and E.-P. Lim, "Comparing Twitter and traditional media using topic models," in *Proc. 33rd Eur. Conf. IR Res. Adv. Inf. Retr. (ECIR)*, 2011, pp. 338–349.
[40] S. Ali. *Express News | Express News Urdu | Latest Urdu News*. Accessed: Oct. 17, 2017. [Online]. Available: https://www.express.pk/
[41] S. A. Hussain. Nawaiwaqt—Latest Urdu News From Pakistan. Islamabad, Lahore, Pakistan. Accessed: Oct. 17, 2017 http://www.nawaiwaqt.com.pk/home
[42] *Latest Breaking News from Pakistan | Pakistan News | Daily Urdu News | Roznama Dunya*. Accessed: Oct. 17, 2017. [Online]. Available: http://dunya.com.pk/
[43] S. Bird and E. Loper, "NLTK: The natural language toolkit," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, pp. 1–4.
[44] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, "Similarity-based machine learning methods for predicting drug-target interactions: A brief review," *Brief. Bioinform.*, vol. 15, no. 5, pp. 734–747, 2013.
[45] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," *Assoc. Comput. Linguistic*, vol. 146, pp. 1301–1309, 2011.
[46] M. Omar, B.-W. On, I. Lee, and G. S. Choi, "LDA topics: Representation and evaluation," *J. Inf. Sci.*, vol. 41, pp. 1–14, Jun. 2015.
[47] S. Zhu, L. Liu, and Y. Wang, "Information retrieval using Hellinger distance and sqrt-cos similarity," in *Proc. 7th Int. Conf. Comput. Sci. Educ. (ICCSE)*, 2012, pp. 925–929.
[48] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. Int. Conf. Database Theory*. Berlin, Germany: Springer, Jan. 2001, pp. 420–434.
[49] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. Workshop New Challenges NLP Fram.*, 2010, pp. 45–50.

WAHEED ANWAR received the master's and M.S. degrees in computer science from The Islamia University of Bahawalpur Punjab, Pakistan, where he is currently pursuing the Ph.D. degree in computer science with the Department of Computer Science and Information Technology (DCS and IT). He has over 10 years of teaching, research, and development experience. He is currently a Lecturer with DCS and IT, The Islamia University of Bahawalpur. He has published three research papers in recent years. His current research interests include text mining, Web mining, machine learning, and deep learning. In addition, he is a Sun Certified Java Programmer SCJP2.

IMRAN SARWAR BAJWA received the Ph.D. degree in computer science from the University of Birmingham, U.K. He is currently an Assistant Professor with the Department of Computer Science and Information Technology, The Islamia University of Bahawalpur. He has very good programming skills in Java and C#. He has published over 130 papers in well-reputed peer-refereed conferences and journals. He has 850+ citations in Google Scholar. In addition, he has 40+ articles in ISI Web of Science, 70+ articles in Scopus, and 300+ citations of his work in Scopus. He has over 15 years of teaching and research experience at various universities of Pakistan, Portugal, and U.K.. He was the Chair of ACM Bahawalpur Chapter from 2016 to 2017.



M. ABBAS CHOUDHARY received the Ph.D. degree in engineering management from The George Washington University, USA. He is currently a Professor and a Rector of the Dadabhoy Institute of Higher Education, Karachi. Previously, he served NAMAL since 2015, first as a Professor and then as the Director. An eminent academician, he has vast experience in the field of higher education management. He has already served two four-year terms as the Vice Chancellor of the University of Engineering and Technology, Taxila, and The Balochistan University of Information Technology, Engineering and Management Sciences, Quetta.

SHABANA RAMZAN received the MCS degree from the University of Engineering and Technology at Lahore, Lahore, Pakistan, in 2004, and the M.S. degree in computer science from the Islamia University of Bahawalpur, Pakistan, in 2014, where she is currently pursuing the Ph.D. degree. She is currently an Assistant Professor with Government Sadiq College Women University, Bahawalpur. Her current research interests include NoSQL databases, NoSQl databases security, cloud networking, and cloud computing. She is a Professional Member of the ACM and the Vice Chair of the ACM Chapter, Bahawalpur.