IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 13, 2021, accepted September 8, 2021, date of publication September 14, 2021,
date of current version September 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3112620

Analyzing LDA and NMF Topic Models for Urdu Tweets via Automatic Labeling

ZOYA[1], SEEMAB LATIF[1], (Senior Member, IEEE), FAISAL SHAFAIT[1,2], AND RABIA LATIF[3]
[1]School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[2]Deep Learning Laboratory, National Center of Artificial Intelligence, Islamabad 44000, Pakistan
[3]College of Computer and Information Science, Prince Sultan University, Riyadh 12435, Saudi Arabia

Corresponding author: Seemab Latif (seemab.latif@seecs.edu.pk)

ABSTRACT The understanding and analyzing of available content on Social media Platforms such as Twitter and Facebook, through various topic modeling methods is not supervised. However, despite several existing conventional techniques, they have had limited success when applied directly for filtering and quick comprehension of short-text contents due to text sparseness and noise. Thus, it always has been challenging to discover reliable latent topics from online discussion texts that prevail with low words co-occurrence and availability of large size social media benchmark datasets, even for resource-rich languages. The existing literature lacks such work for Urdu text to unveil niche topics even with conventional topic models, mainly due to the lack of benchmark datasets, limited availability of pre-processing tools/ algorithms, and time and compute limitations on large-sized datasets. This work presents experiments with multiple approaches of topic modeling like Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet Allocation (LDA), and Non-negative Matrix Factorization (NMF) on 0.8 million Urdu tweets. These tweets are collected through Twitter API by giving various hashtags as a query to avoid dominance of single topic in the dataset. In addition, we have pre-processed the text of the tweets, prepared the three variants of the collected dataset, and extracted multiple features to represent documents on different n-grams. Furthermore, all these techniques are compared and evaluated on the dataset variants, using both qualitative and quantitative measures. We have also demonstrated the results of these approaches through visualization methods, graphs depicting tweets size per topic, word clouds, and hashtags analysis, giving insights about algorithms performances on finalized topics. Observed results reveal that NMF outperformed the techniques with TF-IDF feature vectors in Urdu tweets text, while LDA performed best with merging short-text strategy into long pseudo documents.

INDEX TERMS Natural language processing, short-text topic model, topic evaluation, topic modeling, Urdu text processing.

I. INTRODUCTION

With the prevalence of content sharing platforms, like online forums, macro, and microblogs, social networks, photos and videos sharing websites, people are relatively more familiar with expressing and sharing their opinions on the Internet. The increasing popularity of platforms, like Twitter, leads to large amounts of online discussions every day. As indicated by the 2020 statistics [10], around 340 million people use Twitter every month, 500 million tweets are sent every day and on average around 6,000 tweets are posted every second. This platform permits users to post a short tweet in 34 different languages that include Urdu, Arabic, English, Bengali, Chinese, French, Spanish, including others. However, the massive quantity of Twitter messages makes it impossible to manually process these tweets to determine the subject of public discussions. Topic modeling is an unsupervised machine learning technique that can serve as a useful tool to auto-discover the subject about a document from enormous texts and also can assist people to better understand the most clues and semantic structures. In linguistics, the concept of ''topic'' was initially described as an item about which a sentence is written. When this idea was extended to an entire document, the concept of ''topic'' became a certain distribution of words in a document.

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

Earlier, topic modelling was notably achieved through most popular techniques such as Latent Semantic Analysis (LSA) [9], Probabilistic Latent Semantic Analysis (PLSA) [14], Latent Dirichlet Allocation (LDA) [4] and Non-negative Matrix Factorization (NMF) [18]. Models like LSA, PLSA, and LDA are based on the same basic assumption i.e., each document consists of a mixture of topics and each topic consists of a collection of words. These models split the document-word matrix into the document-topic and topic-word matrices. The distribution of the document-topic matrix represents the degree of each document that belongs to each topic, whereas the topic-word matrix indicates the degree of each word that belongs to each topic. Alternatively, NMF can also be applied to textual data to reveal topical structures. It learns topics by directly decomposing the term-document matrix into two low-rank factor matrices. These types of models have shown outstanding performance for high-dimensional data in dimension reduction and clustering tasks [41]. Conventional topic models have achieved great success in various applications but experienced a large performance degradation over short texts due to a high degree of noise (slang words, emojis, etc.) and data sparsity in social media text. Former techniques deal with the occurrences of words in short documents, as they have a less discriminating role as compared to the lengthy documents. Latter techniques address unproductive jokes, insults, and cursing in tweets, comments, reviews, etc. that are not related to the underlying topics. Hence, document-level word co-occurrence information lacks in the short-text document and the use of informal language causes an increase in dictionary size. Multiple strategies have been adopted in literature to tackle these problems in the short- text including aggregating short text into long-pseudo documents, using auxiliary information like hashtags, time, authors, etc., and self-aggregation strategy that uses prior knowledge to guide conventional models in the form of an external corpus or word embedding, leveraging conversational tree structures and global word co-occurrences. All these strategies are discussed in detail in section II.

The language support provided by Twitter and various other online platforms opens research opportunities in low-resource languages like Urdu. Urdu is the national language of Pakistan. There are approximately 16 million native Urdu speakers, up to 94 million second-language Urdu speakers in Pakistan and more than 300 million speakers are in various parts of the world [8]. Due to advancement in language typing tools for Urdu language, its speakers progressively communicate their opinions on government initiatives, societal problems, religious affairs, and other domains on online platforms including Twitter. However, it has been explored that most of the work done to date in the domain of topic modeling, is relevant to English and other resource-rich languages, only. Urdu is a low-resource language and is different in its syntax and morphological structure from English and many other resource-rich languages. This requires extensive pre-processing efforts and also demands some knowledge about

FIGURE 1. Urdu letter ‘م’ shapes in isolation, start, middle and end of a word.

the language, while handling Urdu text before applying topic models. In Urdu, words are not always separated by whitespaces, so applying an English language tokenizer directly based on whitespaces can create many bad tokens or long phrases that either causes a model to crash while processing the text, or oftentimes ends with poor results. Additionally, Urdu text can be written in various ways like words having the same meanings can be written in various forms e.g. ‘کے لئے’ as ‘کےلئے’. Urdu letters changes shapes depending upon their position in the word; e.g. Urdu letter ‘م’ has different shapes, see figure 1. Moreover, the major hindrances in way of Urdu text processing are that most common text editing software do not support Urdu text, common language-specific resources, such as NLP tools are either limited or unavailable, domain-specific frequent or stopwords lists, lexicons, and benchmark tagged corpora. Thus, without handling these issues, noise and data sparsity problems due to the limited number of co-occurrences can be increased in topic models. Furthermore, existing datasets used for different tasks do not always consist of texts carefully collected by experts, nor are they always appropriate with the model's purpose and desired outcome, small in size, or often biased towards particular contexts. Similarly, evaluation of results is also challenging due to the unavailability of large sized benchmark datasets.

In this connection, this paper aims to devise an analysis of topics people discuss on Twitter in the Urdu language. Due to the lack of a standard benchmark dataset in the Urdu language, we have first collected tweets by leveraging the Twitter API, Tweepy. Since advanced topic modeling techniques have their limitations that either need additional resources like trained word embedding, gold standard datasets, prior knowledge, etc or imposed additional restrictions besides global word co-occurrences, like single topic assumption by losing the flexibility of capturing multiple topics and directly capturing bi-term frequencies with less semantic relatedness, etc. So, different conventional topic modeling algorithms, such as LSA, PLSA, LDA with its extension Hierarchical Dirichlet Processes (HDP) and NMF have been experimented with and compared as an initial step to devise a deeper analysis about these models for Urdu text tweets. We have attempted to established a baseline for advanced techniques to observe their performances in comparison to conventional models. Variants of the collected dataset have been prepared to observe the performance of these models. As tweets come under the short-text document, so we have also experimented with the short text strategy i.e., merged tweets based on long

pseudo documents with hashtag correlation analysis and then applied the LDA model on aggregated Urdu text. The major contributions of our work are as follows:

- Collecting a large number of Urdu tweets for modeling topics and clustering.
- Preparing dataset due to unavailability of any gold-standard dataset in the Urdu language for different experiments.
- Preparing multiple variants of the dataset based on common word co-occurrences.
- Developing an approach for modeling topics based on multiple feature representations with different n-grams extracted from the dataset.
- Analyzing hashtags for abstract comprehension of the topics.
- Comparing, evaluating, and visualizing topic modeling algorithms for Urdu tweets using both qualitative and quantitative methods.

The rest of the paper is organized as follows. Section II presents the literature review of topic modeling in Urdu, English, and other languages. The proposed methodology is illustrated in Section III, while Section IV describes the experimental setup and evaluation of various topic modeling techniques. Results and findings are also discussed in this Section. Section V describes the Evaluation techniques used in this paper with their limitations. Finally, Section VI concludes the paper with future work.

II. LITERATURE REVIEW

In this section, we have discussed the topic modeling techniques for long and short-text and their analysis. We have also discussed the topic modeling techniques for the Urdu language.

A. CONVENTIONAL TOPIC MODELLING FOR LONG TEXT

Topic models have been extensively researched with an aim to discover latent topics and often involve the use of LSA, PLSA, LDA, and NMF to reveal topical structures through considering document-level word co-occurrence patterns. The early LSA follows a matrix factorization approach, constructing a sparse document-term matrix with a row for every document, whereas columns represent all unique terms within the dataset at hand. LSA finds low-dimension representation of documents and words using the Singular Value Decomposition (SVD) technique. However, LSA lacks the differentiation of polysemy words, and the cost of applying SVD is very high. It also fails to update new documents immediately. Afterward, Probabilistic Latent Semantic Analysis uses a probabilistic method, instead of SVD to tackle the problem that exists in LSA. The core idea is to find a probabilistic model P (d, w) with latent topics, such that for any document d and word w, P (d, w) corresponds to an entry in the document-term matrix. Since, PLSA had no prior parameters to the model distribution of words in the corpus, so it resulted in overfitting, given linear increments in the number of documents. Then, LDA was introduced, which allows multiple topics that are represented by multi nominal topic-word distributions for one document. The documents may possess different topic structures that are represented by the document-topic distribution. The usage of Dirichlet priors α and β help to estimate document-topic density and topic-word density, respectively. Thus, LDA overcame the problem parameter estimation process. Nevertheless, LDA also has some limitations that cause it to produce irrelevant topics. Therefore, it has been extended for advanced modeling techniques in terms of supervised models, weighted models, determining the correlations among topics, finding other statistics besides words co-occurrences, removing bag-of-words assumption by considering word order, etc. to tackle its limitations [24]. For example, HDP is one of its extensions that model topics as mixtures of words much like LDA, rather than documents being mixtures of a fixed number of topics. It regulates the number of topics itself during posterior inference. Likewise, some studies are based on replacing uniform term weighting in LDA and aimed at punishing irrelevant words by giving more weight to informative words for topic inference. On other hand, NMF is a factorization-based algorithm that decomposes high-dimensional vectors into a lower-dimensional representation, where we constrain the coefficients of matrices to be non-negative [18]. NMF gives two matrices W and H, where W is the topics found, and H is the coefficients (weights) for those topics.

B. TOPIC MODELLING FOR SHORT TEXT

With the rise of social media in recent years, many studies have specifically covered the contribution of topic models used for online social media networks short texts, like Twitter, Facebook, etc. These researches have demonstrated that although, conventional topic models have achieved great success for the normal text, yet they are not suitable for short text topic modeling due to the high degree of noise and data sparsity in short texts. All of the techniques adopted in literature to tackle such problems have been summarized in Tab 1 alongside their limitations for short-text documents. In this connection, merging short texts into long-pseudo documents by leveraging auxiliary information like author, time, hashtags, location, etc. has been widely adopted to apply standard conventional models on pooled documents. For example, Steinskog *et al.* utilized the idea of pooling schemes by aggregating similar tweets with analysis of hashtags co-occurrences and tweets sharing common author into single long pseudo-document [42]. However, this strategy has limitations, like auxiliary information is not always available, e.g. news title, irrelevant short text can be aggregated, unnecessary hashtags in training data, topics can be biased towards generated pseudo document, and semantically related words may have less or zero co-occurrence. In the meanwhile, researchers introduced another emerging solution to prepare long-pseudo documents without using auxiliary information for short text. Such models utilized a self-aggregation strategy to combine text as a part of their own generative process and provide

informative cross-text word co-occurrences rather than a direct aggregation of short texts. Zuo *et al.* introduced novel probabilistic models called Pseudo-document-based Topic Model (PTM) and Sparsity-enhanced PTM (SPTM) for a short text that was based on self-aggregation strategy to merge short text implicitly into long pseudo-document by utilizing much fewer pseudo-documents [56]. SPTM also used Spike and Slab prior with the purpose of eliminating undesired correlations between pseudo-documents. This approach still could not deal with the data sparsity problem as the clustering method faces the same limited word co-occurrences information. Moreover, time complexity and over-fitting issues have been observed in these models due to an increase in parameters with data size. Several attempts of the like have been made for model improvements and captured corpus-level word co-occurrences to model topics. Some methodologies imposed additional restrictions beside global word co-occurrences that each document consists of a single topic to alleviate the problem of sparsity [37], [54], [56]. However, this assumption becomes too strong in some situations, loses the flexibility of capturing multiple topics in a document, causes over-fitting, and fails to resolve data sparsity issues. Some researchers directly captured disordered word-pairs, co-occurred in the same document as bi-terms, and modeled topics on them by assuming that these bi-terms are generated from the same topic-word distribution [7], [15], [50]. Another line of research focused on enriching prior knowledge into topic models for their guidance towards topic inference by utilizing some external corpus. This helps them perform focused analysis towards finding interesting topics on a particular feature. For example, Wang *et al.* have proposed a novel idea of Targeted Topic Model (TTM) to enable focused analysis on any specific aspect instead of performing a full analysis of the entire dataset for discovering inherent topics [46]. Miller *et al.* have proposed an informed LDA topic model to noisily label the documents without any supervision in order to train classifiers with labeled instances [28]. Crowd-sourced knowledge contained in Wikipedia was used in this context as informed prior to LDA. Many studies are based on trained word embedding concepts used as prior knowledge to leverage internal semantic information as well beyond simple word co-occurrence information. Shi *et al.* have proposed a Semantics-assisted Non-negative Matrix Factorization (SeaNMF) model based on a block coordinate descent algorithm to discover topics from microblog short texts [41]. Semantic correlations of contextual words were incorporated into the model using the Skip-gram view of the corpus. Gupta *et al.* addressed the problem of less context and data sparseness in short text for modeling topics and proposed a model named iDocNADE [12]. The proposed model was an extension of the DocNADE model that incorporated full contextual information in terms of word embedding. Furthermore, there are some models that jointly learn word embedding and infer topics on the same corpus in a unified manner, instead of using existing word embedding trained on some external corpus. Shai *et al.* proposed a framework, the Skip-gram Topical word Embedding (STE) model, which can learn word embedding and latent topics with mutual interaction between two paradigms [40]. The framework combines the Expectation-Maximization (EM) method with the negative sampling scheme to improve the quality of discovered topics. Lu *et al.* have proposed a model named RNN-IDF based Biterm Short-text Topic Model (RIBS-TM) [25]. This model makes use of Recurrent Neural Network (RNN) for relationship learning between words and inverse document frequency (IDF) for filtering high-frequency words. Preparing word embedding in the generation process of topic models can relieve the curse of data sparsity but can suffer from unacceptable time complexity. Little work has been conducted on using conversational tree structures as prior information or exploiting discussion threads to organize post messages in form of parent-child relation to discover topics.

C. TOPIC MODELLING FOR URDU LANGUAGE

As far as the Urdu language is concerned, a limited number of studies have addressed the topic modeling domain especially for self-generated short-text corpora, and just the LDA model has been experimented with for modeling topics in all of the studies [31], [32], [39]. Furthermore, their experimental results in terms of topics evaluation have not been explained well. Comparison with other standard topic models has not been discussed at all with their coherence scores or with some downstream tasks. For example, Munir *et al.* performed experiments for comparison of various topic modeling approaches, including LSI, LDA, and HDP on Urdu poetry and news headlines texts [31]. They concluded that LDA outperformed on Urdu news headlines text but all of them were not suitable for Urdu poetry text. However, topics that have been compared by pyLDAvis visualization technique, is deemed not to be an appropriate approach for topics comparison due to no relation among information obtained through its different projections (e.g., t-SNE, pcoa, mmds) [6], [26]. Moreover, it is quite difficult to determine the relative size of circles and the distance between them to analyze importance of the topic. Likewise, Nasim and Haider [32] experimented with clustering approaches including K-means, Bisecting K-Means, and Affinity Propagation algorithms on various features (e.g., TF-IDF and embedding) extracted from tweets dataset and then compared with the LDA topic model on its default feature vector. The main focus of the study was on clustering algorithms and other topic models including LDA. It has been concluded that TF- IDF feature vector combined with the K-means clustering algorithm outperformed the adopted techniques and LDA performance is not up to the mark with a short text document. Again, there is no discussion about topic coherence, their merging the tweets with hashtags, and hashtag pooling technique results. Shakeel *et al.* proposed a framework for modeling topics, specifically for Urdu text called Topic Model for Urdu (ULDA) that combined preprocessing techniques and LDA model with Gibbs sampling

TABLE 1. Comparative analysis of techniques for short text topic modelling.

Strategy	Techniques	Used by	Datasets	Limitations	Urdu
Conventional / local word co-occurrence based	LSA, PLSA, LDA, NMF, ULDA	[34], [56], [41], [25], [46], [47], [51]	20-Newsgroups (online posts on 20 topics), TagMyNews(news), Sanders Twitter corpus, News titles, DBLP (paper titles), Questions, Tweets, yahoo.Ans, Product reviews, Urdu news headlines, and articles, Urdu Textbooks, Urdu miscellaneous poetry, Urdu Wikipedia dump, Urdu tweets	Severe data sparsity issue due to document-level word co-occurrence. Ignore metadata information in semi-structured texts like tweets. Bag-of-words assumption neglects word order. No prior knowledge and cannot self regulate topics. Consider no relation among topics. Uniform term weighing inside topic. Fail to leverage semantic information between words	yes
Term weighing based	TWLDA, WTM, BWTM, CEW	[51], [19], [21]	Amazon reviews, 20-Newsgroups(online posts on 20 topics), TREC (questions),Snippets(web searches), Biomedical, StackOverFlow, WebKB (web pages),Reuters (news documents) and ohsumed (abstracts)	Static weighing (e.g. TF-IDF) and high computation cost, punish just meaningless words like stopwords, ignore topic discriminatory words by focusing just on documents	No
Long pseudo-document using auxiliary information	LDA, ATM, HGTM	[27], [44], [36], [2], [47], [42]	Tweets	Highly data dependent, Semantically related words may have zero or less probability, Discovered topics may be biased towards pseudo document, Irrelevant text may be aggregated e.g. unnecessary hashtags in the data	yes
Self-aggregation based	SATM, PTM, SPTM, DSTM	[37], [56], [23]	DBLP (paper titles), 20-Newsgroup, Tweets, Online news, Questions, NIPS (conference paper information), Yahoo! Q&A	Clustering method faces the same problem of limited word co-occurrences to corpus itself, Parameters increases with data size causing over-fitting and time complexity problems, Unable to fully capture the semantic relatedness among words, No prior knowledge to ensure the quality of aggregated text	No
Single topic per document assumption / Global word co-occurrence based	Twitter-LDA, MU, PTM, SATM, DMM and its variants (GPU-DMM, GS-DMM, LF-DMM), BTM, BPDTM, PMI--BTM, WNTM, GPU-PDMM	[54], [56], [37], [35], [20], [52], [34], [50], [15], [7], [57], [20]	Tweets, News, DBLP, Questions, NIPS, Yahoo! Q&A, 20-Newsgroups, TagMyNews(news), Snippets, BaiduQA, Sanders Twitter corpus	Single topic assumption loses the flexibility of capturing multiple topics in document and causes over-fitting, Simple bi-term frequency causes topics containing common words, Limited corpus-level word co-occurrence information can't alleviate data sparsity issue completely, Fail to utilize external semantic information to deal with less context	No
Prior Knowledge based/ Embedding based	TTM, InformedLDA, RLTM-SK, LF-LDA, LF-DMM, WEI-FTM, SeaNMF, SSeaNMF, IATM, CGTM, iDocNADE, GPU-DMM, ETM, STE, GLDA, MetaLDA, GPU-PDMM	[46], [28], [48], [34], [53], [41], [13], [49], [12], [20]	Amazon product reviews, 20-Newsgroups, TagMyNews, Sanders Twitter corpus, Reuters, KOS, Snippets, TagMyNews, Tweets, DBLP, Yahoo! Q&A, News, BaiduQA,	High dependency on the quality and availability of external corpus in same domain, Domain difference may cause noise and bias in topics, Benchmark large size corpus in same domain is not always available	No
Topical Embedding based	STE, RIBS-TM, VTMRL	[40], [25], [11]	Chinese online questions and news, 20-newsgroup, NIPS	Semantic information is limited to target corpus itself, Computationally expensive	No
Conversational Tree Structure based	LeadLDA, HDP, CSATM, IATM	[20], [45], [43], [13]	micro-blogs conversation trees, Reddit(online messages)	May have time complexity for computing root to leaf path model robustness issues, Unable to build hierarchical structure of topics, Fail to discover topics with multi-conversation tree	No

to boost the performance of information retrieval (IR) system [39]. The proposed model was implemented on three self-generated corpora with relatively small sizes, including normal and short text. Nevertheless, the performance of the model specifically on short text, their coherence scores, and any abstract knowledge about several topics inside the corpus has not been discussed. In addition, Anwar *et al.* proposed LDA instance-based and LDA profile-based approaches to classify authors of Urdu texts in an unsupervised manner [3]. The presented work dealt with diversity in writing styles and ambiguity of normal size text inherent in the Urdu language. Recent research in the domain of topic modeling for the Urdu language has indicated that this field has not been explored well, despite progressive research in other resource-rich languages. Even basic conventional models and strategies to deal with short text have not been experimented with and evaluated properly to see their overall performance; besides LDA to some extent with little discussion about results. Therefore, our work is the first step with detailed quantitative and qualitative evaluation results on these models with Urdu text on a comparatively very large tweets dataset. A comparison of approaches applied to the Urdu language with other short text strategies has been conducted in Table 1.

III. METHODOLOGY

The proposed methodology is divided into three major steps, as shown in figure 2. Initially, we have observed top Twitter trends irrespective of the specific domain by using Twitter API and filtered English or other language tweets, retweets as well as truncated tweets. Extracted tweets are stored in a database, where duplicate tweets are filtered and further preprocessing and hashtags correlation analysis is carried out. Different variants of the dataset are prepared based on their analysis and their multiple features are extracted for thorough experimentation. Further details of these steps are discussed in the following steps:

A. DATASET COLLECTION AND CONSTRUCTION

We have collected Twitter data by leveraging Python Twitter mining and authentication API, tweepy. The period for collecting tweets spans from April 2020 till August 2020. Twitter API allows the extraction of tweets for the last 7 days using a single hashtag as a query term with some simple operators. Moreover, there is a limit of searching tweets in standard API i.e., 180 requests per 15 minutes window for a single authenticated user on Twitter, and a maximum of 100 tweets can be scrapped per query request. We have collected approximately 5 million tweets in 5 months by querying the Twitter track multiple times. Our query terms consisted of top Twitter hashtag trends in Pakistan observed on daily basis. We drew out Urdu tweets and also filtered out truncated tweets and retweets. Furthermore, our dataset is based on generic query terms irrespective of specific hashtags category, thus covering almost all topics discussed in tweets for the given period. For conducting multiple experiments, we have generated multiple versions of the datasets, as given in Table 2. These versions are discussed below:

1) GENERIC DATASET (DATASET-A)

This dataset consists of 0.89 million pre-processed tweets collected using hashtags as a generic query term. Further, two forms of this dataset were prepared with little difference.

2) BI-GRAM READY GENERIC DATASET (DATASET-B)

Frequent co-occurrences of words were observed in the dataset and if two words appear at least five times together, they were treated as a single word by replacing space with an underscore, for example, عمران خان e.g. as عمران_خان (Imran Khan as Imran_Khan).

3) BI-GRAM AND TRIGRAM READY GENERIC DATASET (DATASET-C)

Frequent co-occurrences of words were observed in the dataset and if three words appear at least five times together, they were treated as a single word by replacing spaces with underscores, for example, وزیراعظم عمران خان as وزیراعظم_عمران_خان (PrimeMinister Imran Khan as PrimeMinister_Imran_Khan).

4) HASHTAGS ANALYSED DATASET (DATASET-H)

A single tweet may contain several related hashtags that give an idea about subject discourse in a tweet. Hashtag correlation analysis can help in estimating the number of topics discussed in tweets. In this regard, we have analyzed the frequency of each hashtag and selected the top frequent hashtags that appeared more than 3000 times in the tweets dataset. Afterward, we tried to understand the strength of relationships among them to combine related hashtags into a single group. Tweets that come under one group can be merged into one long-pseudo document. Hence, we computed Pearson's coefficient using equation (1):

$$p(X, Y) = \frac{COV(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

where COV(X, Y) is a covariance between X and Y to analyze the pairwise relationships between hashtags. We have considered only the positive relationships between X and Y by indicating threshold value 0.01, thus ignoring the negative correlations to filter irrelevant hashtags. Hashtag correlations are given in figure 3. The intensity of the blue color indicates a strong positive correlation among hashtag pairs. We have extracted unique and positively correlated pairs, grouped them by taking common hashtag as a key among these pairs, for example, (#ImranKhan, #PTIGovernment) and (#ImranKhan, #LockdownPakistan) into a single group as (#ImranKhan, #PTIGovernment, #LockdownPakistan).

We constructed 38 correlated hashtags groups this way, as well as, long pseudo-documents belonging to each particular correlated hashtags group. Tweets consisting of a minimum of three related hashtags in a hashtag group

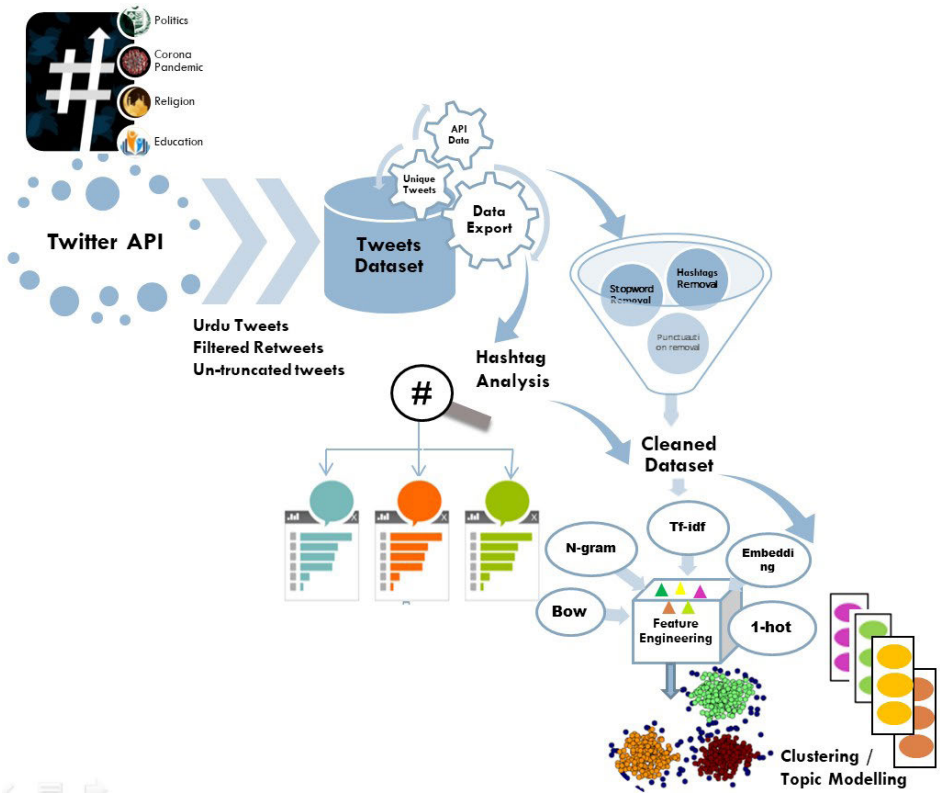

FIGURE 2. Process pipeline of methodology.

TABLE 2. Outline of datasets, feature representations, topic models and evaluation methods used in study.

Datasets	**Dataset-A**: Generic dataset Size = 892,749 pre-processed tweets Number of words = 16,748,557 *Modifications of Dataset-A* **Dataset-B**: Bi-gram ready generic dataset Number of words = 13,588,118 **Dataset-C**: Tri-gram ready generic dataset Number of words = 11,870,269 **Dataset-H**: Correlated hashtags merged tweets 38 long-pseudo documents
Feature Vectors	**F1**: BOW N-grams: uni-gram and bi-gram **F2**: TF-IDF N-grams: uni-gram and bi-gram
Topic models	LSA, PLSA, LDA, NMF
Quantitative Evaluation	Coherence Measures (CM): C_v, C_uci, C_umass, C_npmi, Perplexity, Precision
Qualitative Evaluation	3 human judgments with Cohen's Kappa agreement score

were merged and treated as a single long pseudo-document. We have also removed duplicate tweets to overcome the poor performance of generative topic models on short-text documents. Preparing pooled documents with this strategy help overcoming its shortcomings, like avoiding irrelevant text to be gathered under specific hashtags group. Moreover, hashtags are collected generically based on top Twitter trends, so it avoids generating biased topics towards a particular domain after doing their analysis and grouping related hashtags in one domain. Besides, it will help analyzing the extent of improvement in the results of topic models by overcoming the limitations of short-text documents (like sparsity issues) for Urdu language text. The process followed to generate Dataset-H is given in figure 4.

B. TWEETS PRE-PROCESSING

The acquired tweets from Twitter API consist of many irrelevant features that are not required for detecting topics. We are considering the tweets in the Urdu language, which also require pre-processing for the removal of unnecessary and redundant features. Pre-processing steps performed on tweets are:

1) DUPLICATE TWEETS REMOVAL

A single tweet may consist of many related hashtags for example, '#Covid_19', '#Coronavirus', '#lockdown' etc. By using these hashtags as a single query term may extract one tweet multiple times. So, we kept only the single instance of each duplicated tweet in our dataset.

2) NULL VALUES REMOVAL

Some extracted tweets from the API may contain unnecessary data or no text, so we removed such tweets from our dataset.

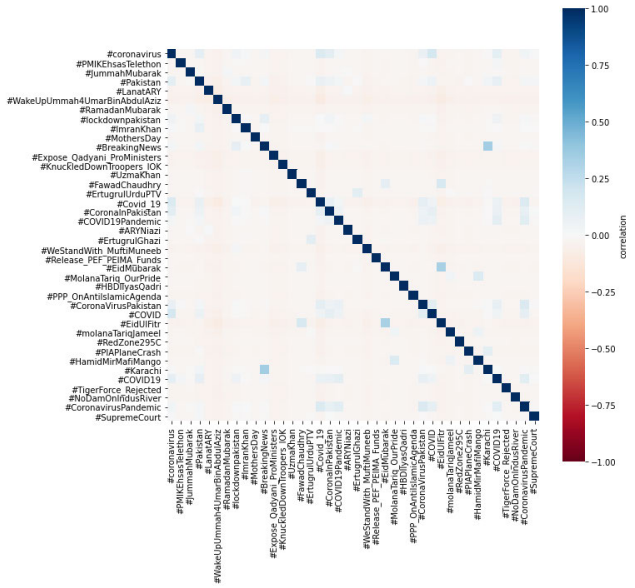

FIGURE 3. Hashtags correlation analysis.

3) NORMALIZATION

By Normalization of Urdu text, we brought all text under the specific range of UTF-8 encoding of Urdu language. This also replaced the Arabic characters with relevant Urdu characters. Moreover, we have observed multiple forms of common words and standardize them in a single form. e.g. 'اسکا' as 'اس کا'.

4) STOP WORDS REMOVAL

Stop words are normally the high-frequency words so we removed the top frequent words in our dataset by observing a specific threshold value, as well as other stop words common in the Urdu language especially in tweets text.

5) PUNCTUATION, DIACRITICS AND EXTRA WHITE SPACES REMOVAL

We removed punctuation, like ” ” ”!”%&’“‘”’()*+, − <=> ; : /.@[]‾‘{}|?\ diacritics and extra white spaces from tweets text. As diacritics are only useful for correct pronunciation of words and single white space between words is sufficient.

6) HASHTAGS, URLS AND MENTIONS REMOVAL

We cleaned tweets text from hashtags, URLs, and mentions (@) as these were irrelevant in the text being prepared for topic modeling.

7) EMOJIS REMOVAL

Emojis are used to express emotions as they have no use in modeling topics, so we removed them from the data.

All above mentioned pre-processing steps brought our dataset down to 0.89 million unique tweets. Tweet pre-processing steps are shown in figure 5.

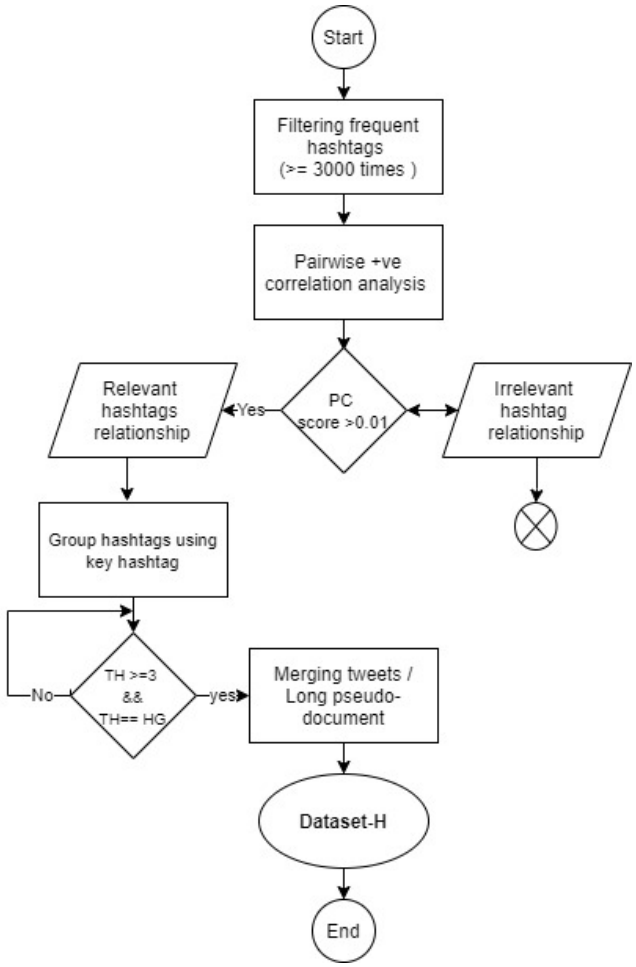

FIGURE 4. Dataset-H construction process (TH denotes single tweet hashtags, HG individual hashtag group and PC Pearson's coefficient score).

C. FEATURE EXTRACTION

Feature extraction plays a crucial role to discover the significant patterns to acquire information. The extracted features are the distribution of different numbers in vector form. These values are known as weights to represent words in some documents. Various techniques to find the values of these weights include: Bag of Words (BOW) vectorization, Term Frequency (TF) vectorization, Term Frequency - Inverse Document Frequency (TF-IDF), context window-based word embeddings, document embeddings, etc. In this research, we have experimented with TF-IDF and BOW vectorizations and built metrics consisting of uni-grams and bi-grams on the generic dataset and its variants as well.

IV. EXPERIMENTATION AND RESULTS

In this section, we have experimented with the most widely used topic models on our Urdu text tweets datasets, feature vectors, evaluation measures, and advances to handle short-text documents, like an aggregated dataset to prepare a long pseudo document. A detailed flow chart of the experiments is given in figure 6. In this regard, we have selected LDA,

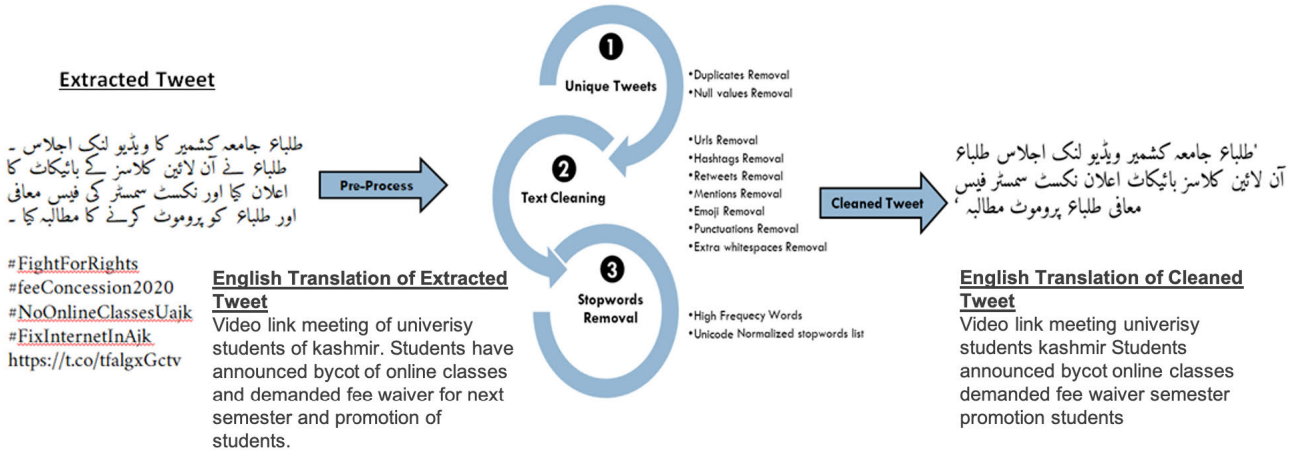

FIGURE 5. Tweet pre-processing.

NMF, LSA, pLSA, and HDP models for exploring relevant topics in our tweets' dataset. All these models except HDP required selecting the number of topics beforehand, so we randomly chose topic counts for this model. We needed to decide an appropriate number of topics, as well as the best model for our Urdu tweets dataset to gain quality results. It is important to look at the highly weighted keywords selected by the model and their semantic relation with each other to identify any topic. Different coherence measures are used to interpret a set of words fitting together with some semantic relationship. These measures can be applied to compare the models by their average coherence score on multiple topics. The best performing model with high coherence score on some specific topic count can be chosen as a final model with that topic value.

A. COHERENCE MEASURE

Coherence measures evaluate a single topic by measuring the degree of semantic similarity between high-scoring words inside a topic. These measures help distinguish between topics that are semantically interpretable from other topics that are artifacts of statistical inference. We have employed four topic coherence metrics in our study to assess the models' performance, including *C_v*, *C_uci*, *C_umass*, and *C_npmi*. *C_v* evaluates based on a sliding window, a one-set segmentation of the top words, and an indirect confirmation measure that uses Normalized Point-wise Mutual Information (NPMI) and the cosine similarity. This measure has been proven particularly appropriate to evaluate the quality of topics based on a large-scale empirical comparison with other widely used topic coherence measures and provides scores closest to human evaluation [38]. *C_uci* is based on a sliding window and computes the Point-wise Mutual Information (PMI) score of all word pairs in the top words list [33]. *C_umass* defines score based on document co-occurrence counts, a one-preceding segmentation, and a logarithmic conditional probability as confirmation measure [29]. *C_npmi* is an advanced version of the *C_uci* that uses NPMI [5].

B. MODEL SELECTION

Due to LDA's popularity in literature, we selected it to conduct the experiments to find out its effectiveness on 45 topics using four coherence measures with step size 5, as shown in figure 7 (i). Since topics count from 5 to 10 was best in all four coherence measures in LDA, so we decided to experiment further to determine the exact number of topics that are best between 5 and 10 with step size 1 (see figure 7 (ii)). Figure 7 (ii) results showed that the coherence score for topic 9 was good in all coherence measures.

Unfortunately, it was not sufficient to cover all dataset topics with significant results. Through hashtags analysis, we had some idea about prevailing topics, but predicted results showed that some topics were missed. Even though we experimented with uni-grams and bi-grams of BOW and TF-IDF feature vectors for LDA and NMF respectively. Uni-gram topics made no sense about the label, see figure 9 and Table 4 for topic 3 and 5. Bi-gram results were somehow better than uni-gram still, topics were more general and overlapping with each other (see figure 9 Topic 7). It was hard to decide an exact subject about topic 7. Analysis of results given in Table 3 with respect to coherence measures is given below:

1) *C_v*: NMF proved the best model among all other models on our Dataset-A. LDA and PLSA were quite close to each other in *C_v* coherence score.
2) *C_umass*: PLSA performed best in comparison to other models. While, LDA and NMF performed equally well and there was a minor difference between coherence scores of LDA, NMF and PLSA.
3) *C_uci*: NMF and LSA performed equally well in their coherence scores.

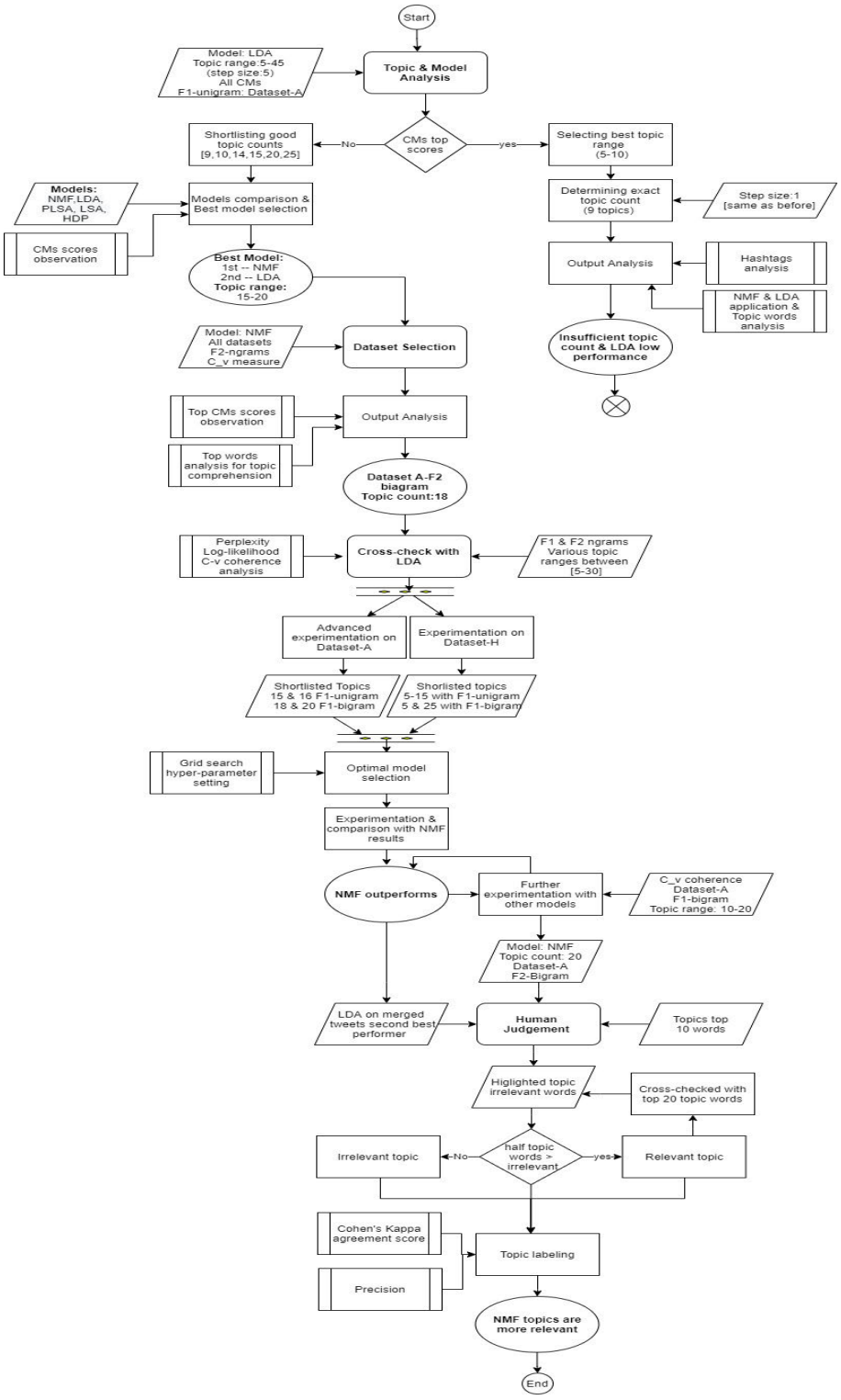

FIGURE 6. Experimentation flow chart.

4) *C_npmi*: NMF achieved the best score against other models. LDA and PLSA were somewhat close to each other.

Afterward, by examining the number of topics, where LDA performed well in all the coherence measures, we choose topics 9, 10, 14, 15, 20, and 25 as topic counts and compared

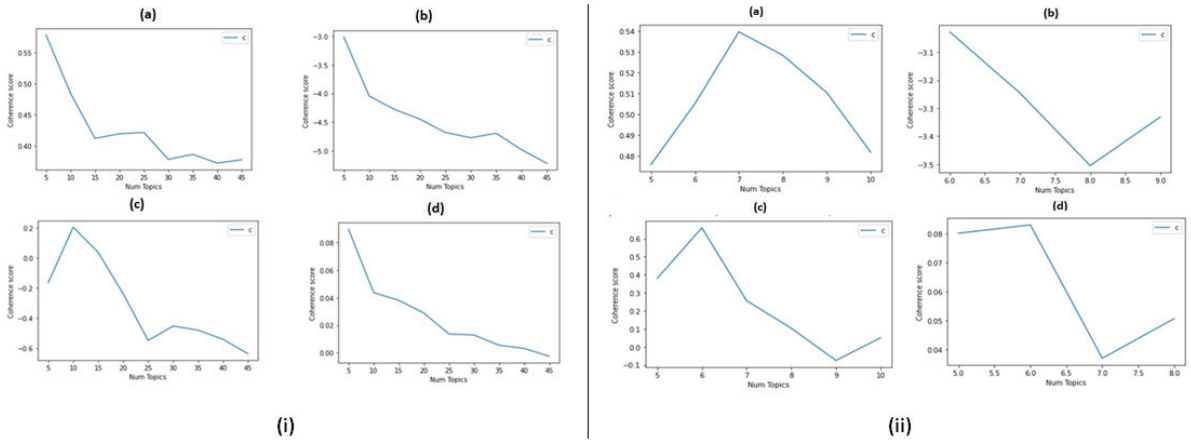

FIGURE 7. Coherence measures with LDA (a) *C_v* (b) *C_umass* (c) *C_uci*, (d) *C_npmi*. (i) shows 45 topics with step size 5 and (ii) shows 10 topics with step size 1.

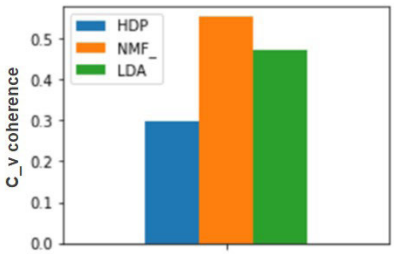

FIGURE 8. Comparison of HDP with LDA & NMF.

models for various coherence measures on these topic counts as shown in figure 10 and Table 3.

Consequently, we kept up with NMF and LDA for further experimentation on the dataset to explore enhanced quality topics. We picked NMF for its performance in contrast to other models and LDA due to its affinity with NMF in the achievement of coherence scores. Additionally, we applied the HDP model that automatically divides the dataset into relevant topics without pre-specifying the topic count. Thus, it generated 19 topics on our Dataset-A, which we compared with LDA and NMF concerning *C_v* measure only on 19 topics. Figure 8 showed NMF as the best model and LDA ranked as the second-best model among three models.

C. TOPIC SELECTION

We have investigated n-gram features to model topics with NMF and LDA on all variants of tweets datasets. In this connection, instead of choosing just selected topics count, we tried looking in detail at the selection of best topic count without any big interval between topic series. Previously, all coherence measures were calculated on a selected number of topics and only uni-gram features were being considered on Dataset-A. Now, we computed *C_v* coherence for NMF and LDA models with uni-gram and bi-gram TF-IDF and BOW feature vectors, respectively, on multiple topics with little or no gap between topics count. Firstly, we experimented with NMF on our tweets datasets for recognizing appropriate topics count. Table 5 showed that NMF performed well on Dataset-A rather on the rest of the datasets and the highest coherence score was achieved for Dataset-A on topic count 18 for uni-gram vectors also shown in figure 11 (i)a. While, figure 11 (i)b and figure 11 (i)c showed that it reached to the highest scores for topic counts 16 and 20 on the other two variants of the dataset, yet their coherence scores were quite low as compared to Dataset-A. We further experimented with NMF with a topic range between 22-30 beyond topic count 20 to investigate the best number of topics with *C_v* coherence scores. Figure 12(a) showed the best coherence score was found on topic count 24 and there was no major difference in performances for bi-grams, figure 12(b).

We applied LDA on Dataset-A and computed *C_v* coherence with the same number of topics as for NMF. We also got the benefit of LDA likelihood and perplexity measures [4] with an intention to discover more appropriate results. It performed well on topic count 16 in terms of *C_v* coherence score and 15 for perplexity and likelihood with minor difference in coherence values as shown in figure 13 and Table 5. Since, LDA overall coherence was lower than NMF results on Dataset-A, so we did not experiment any further for LDA on other variants of the dataset and also based on previously learned experiences (see Table 3), where no model exceeded NMF's *C_v* coherence score.

Additionally, we tried to explore results with bi-gram feature vectors between NMF and LDA on multiple topics of datasets, shown in Table 6. Again, NMF performed better on Dataset-A, also shown in figure 11. It also performed well on Dataset-C but the coherence score on topic 10 is highest as compared to all other topics scores of Dataset-C. Moreover, increasing the number of topics causes low coherence in the case of Dataset-B, and we knew by hashtag analysis that our dataset must contain at least 14 topics, while any topics below this count were insufficient to deal with all prevailing topics.

TABLE 3. Models coherence measures scores with uni-grams on multiple topics.

Model	LDA				NMF				LSA				PLSA			
Feature	BOW uni-gram				TF-IDF uni-gram				BOW uni-gram				BOW uni-gram			
Topic	*C_v*	*C_umass*	*C_uci*	*C_npmi*	*C_v*	*C_umass*	*C_uci*	*C_npmi*	*C_v*	*C_umass*	*C_uci*	*C_npmi*	*C_v*	*C_umass*	*C_uci*	*C_npmi*
9	0.46	-2.84	0.31	0.05	**0.55**	-2.87	**0.49**	**0.07**	0.37	-3.39	-0.68	-0.003	0.49	**-2.82**	0.35	0.05
10	0.47	-2.88	0.31	0.05	**0.55**	-2.86	**0.51**	**0.07**	0.37	-3.35	-0.62	-0.002	0.49	**-2.85**	0.34	0.05
14	0.51	**-2.94**	0.41	0.06	**0.55**	-2.95	**0.53**	**0.07**	0.36	-3.23	-0.52	-0.01	0.49	-2.96	0.30	0.05
15	0.51	**-2.95**	0.37	0.05	**0.57**	**-2.95**	**0.59**	**0.08**	0.37	-3.20	-0.44	-0.002	0.49	-2.99	0.31	0.05
20	0.51	-3.17	0.32	0.06	**0.56**	-3.11	**0.58**	**0.08**	0.37	-3.18	-0.35	0.002	0.51	**-3.07**	0.38	0.05
25	0.52	-3.20	0.38	0.06	**0.56**	-3.20	**0.58**	**0.08**	0.38	-3.15	-0.30	0.002	0.52	**-3.14**	0.43	0.06

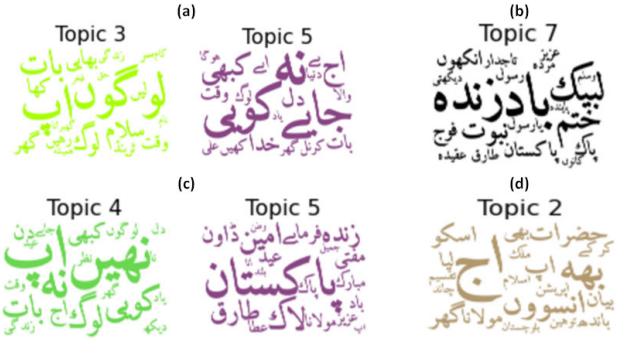

FIGURE 9. Topics word cloud (a) NMF uni-gram (b) NMF bi-gram (c) LDA uni-gram (d) LDA bi-gram.

TABLE 4. Topics word cloud english translation (top 10 words).

Topic 3 (a)	People, talk, you, brother, home, time, salam, life, eat, take
Topic 5(a)	Any, ever, go, not, heart, God, today, talk, colonel, time
Topic 7(b)	Labaik, end, zindabad, prophethood, army, eyes, Pakistan, Tariq, belief, death
Topic 4(c)	You, no, any, people, talk, today, day, life, eid, time
Topic 5(c)	Pakistan, lockdown, molana, Tariq, bestow, Mubarak, eid, mufti, alive, Aameen
Topic 2 (d)	Today, him, country, molana, tear, home, people, moon, divide, tie

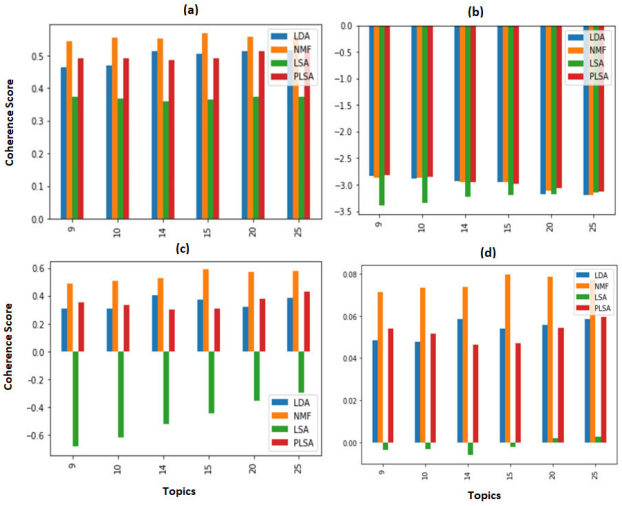

FIGURE 10. Models' coherence measures with uni-gram (a) *C_v* (b) *C_umass* (c) *C_uci* (d) *C_npmi*.

On the other hand, Table 6 showed LDA coherence scores were quite low on Dataset-A. Later we applied the Grid search technique to find the leading LDA model with the best hyperparameter settings for bi-gram feature vectors, due to its low coherence values on all topic counts. Since, LDA was compute expensive so we selected only 2 topics, 18 and 20, where perplexity, likelihood, and coherence scores were comparatively good to learn about best learning decay by the model. Analysis showed that 18 topics were best with 0.7 learning decay. Despite using the best learning decay found by the Grid search algorithm, discovered topics were semantically less coherent than NMF. Here once again we dropped the notion of further experimentation with LDA on other variants of the dataset. So far, the outcomes revealed that the NMF model was verified to be the best model with 20 topics on Dataset-A. Lastly, we compared all models with NMF for *C_v* coherence scores with bi-gram vectorization to get improved justification of achieved results. NMF demonstrated the best score amongst other models on all topic counts as presented in figure 14. Finally, results achieved by NMF on topic count 20 can be seen in Table 7, where coherence score for the individual topic has been given and also labels have been assigned to each topic by looking at the top 10 high ranked words. Only topics 4, 11, 15, and 19 have low coherence scores, i.e. less than 0.5.

1) TOPIC MODELLING ON LONG PSEUDO DOCUMENTS

We performed LDA on merged tweets of hashtags analyzed Dataset-H with topic range 5-30 at step size 5 and measured *C_v* coherence on each topic count for uni-gram as well as bi-gram feature vectors. Table 5 revealed better coherence scores than previously performed experiments, in which every tweet was treated as an individual document. Furthermore, we also calculated perplexity and log-likelihood on each topic count of the aforementioned topic range and the best scores are displayed in bold figures. The best coherence was found on topic count 5, while perplexity and log-likelihood were best on topic count 15 for uni-gram vectors. We analyzed the top 10 words of each individual topic along with their coherence values. Results given in Table 8 showed that topics were few and more general with topic count 5. On other hand, mixed and unknown types in topic labels were more with topic count 15. Hence, we decided to reduce the count from 15 to 10, a second-best option in terms of coherence score. We achieved a better outcome as

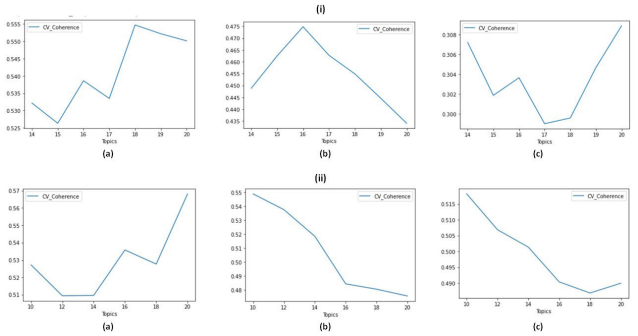

FIGURE 11. Datasets NMF *C_v* coherences (top 20 words) with TF-IDF: (i) uni-gram (ii) bi-gram (a) Dataset-A (b) Dataset-B (c) Dataset-C.

TABLE 5. NMF & LDA uni-gram *C_v* coherence scores.

NMF				LDA		
Features	TF-IDF uni-gram			BOW uni-gram		
Dataset	A	B	C	A		
Topic	C_v Coherence			C_v	Log_Likelihood	Perplexity
15	0.53	0.46	0.30	0.47	**-92739091.71**	**3883.23**
16	0.54	**0.47**	0.30	**0.48**	-92755941.36	3889.06
17	0.53	0.46	0.30	0.45	-92750767.50	3887.27
18	**0.55**	0.45	0.30	0.45	-92826886.24	3913.73
19	0.55	0.44	0.30	0.47	-92830563.98	3915.01
20	0.55	0.43	**0.31**	0.47	-92777080.72	3896.39

TABLE 6. NMF & LDA bi-gram *C_v* coherence scores.

NMF				LDA		
Features	TF-IDF bi-gram			BOW bi-gram		
Dataset	A	B	C	A		
Topic	C_v Coherences			C_v	Log_likelihood	Perplexity
10	0.53	0.40	**0.52**	0.32	-94148675.18	63231.68
12	0.51	**0.42**	0.51	0.34	-94144297.67	63199.18
14	0.51	0.41	0.50	0.34	-94077206.03	62703.28
16	0.54	0.42	0.49	0.34	-94031690.70	62369.07
18	0.53	0.40	0.49	**0.35**	-93957374.74	61827.22
20	**0.57**	0.40	0.49	0.34	**-93938290.07**	**61688.82**

compared to the previous two choices about number of topics. Moving forward with our experiments on bi-gram feature vectors, we examined less coherence score than uni-gram feature vectors but still better than LDA results on bi-gram vectors for Dataset-A (Table 8 and 9). Here we found that 25 is the best number of topics in terms of coherence score and 5 topics in terms of perplexity and log-likelihood. Again 5 topics explored by LDA were more like uni-gram vector results, while topics with count 25 contained many irrelevant topics. Based on previous experience and less difference in coherence score of topic count 10, we also investigated results of LDA with bi-gram vectors on topic count 10 and found unsatisfactory results of less coherent topics.

D. PARAMETERS SETTING

For the settings of NMF, we have used Singular Value Decomposition (SVD) based initialization instead of random seed in order to get more reliable results. In all topic models, we removed the words that exist in less than 10 documents and more than 80% documents. The top 20 high probability words were used to calculate the average coherence with all

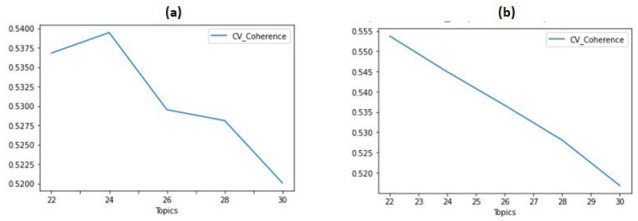

FIGURE 12. Dataset-A: NMF *C_v* coherences (top 20 words) with TF-IDF: (a) uni-gram (b) bi-gram.

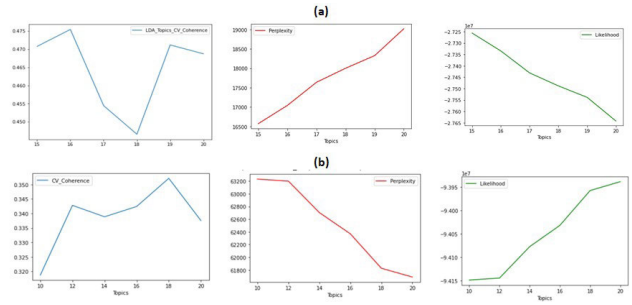

FIGURE 13. Dataset-A: LDA evaluation (a) *C_v* coherence on uni-gram (b) *C_v* coherence on bi-gram with Perplexity & Log-likelihood.

coherence measures for all models in the selection process; whereas, the top 10 words coherence score of each topic has been computed for our final selected model. $\alpha = 0.7$ for LDA has been estimated using the Greedy search algorithm. All other parameters remain the same as per default settings of the Sklearn library in all topic models.

V. EVALUATION MEASURES AND LIMITATIONS

Evaluating a topic model becomes a complicated task in the absence of a gold standard corpus with authenticated topic list, to which results could be compared and measured with well-established tools available to the NLP community. Despite the common use of various topic modeling approaches, it is really hard to evaluate for the right number of topics with a lack of underlying knowledge about actual true labels. Therefore, it is worth emphasizing that evaluation is just a guide to select a better model. Strong hypotheses about the relations between documents should be adopted among a population of topic models, and with differing numbers of topics. Some previous work has been done on evaluating topic modeling results, but they do not directly address the issue of the correct set of topics. Topic evaluation approaches can be grouped into four categories, like Eyeballing models (Termite, pyLDAvis), intrinsic evaluation metrics, human judgments, and extrinsic evaluation metrics. Intrinsic evaluation metrics do not require ground truth labeled datasets, like coherence measures, perplexity, likelihood, etc., while classification techniques come under the umbrella of extrinsic evaluation measures requiring labeled datasets, like Accuracy, Precision, Recall and F1-measure [1], [22]. Furthermore, topic evaluation techniques have their own limitations, which cannot be avoided [6], [16], [26], [38].

TABLE 7. NMF produced topics on dataset-A.

Topics	CV	Top 10 Words	Labels
1	0.82	علیہ والہ، والد وسلم، صلی علیہ، اللہ علیہ، صلی اللہ، رسول صلی، اپ صلیٰ وسلم صلیٰ محمد صلیٰ صلی علیہ (Words of Darood Shareef- a specific saying to compliment Prophet Muhammad (PBUH))	Prophet Mohammad (SAW)
2	**0.96**	رسالت پہ، جان رحمت، نبوت پہ، سے لاکھوں، سےحد، یہ سے، بات میرے، واہ بات، سلام واہ، میرے رضا (on apostleship, Jaan-e-rehmat, on prophethood, aye millions, immensely, this, my talk, wah talk, greetings wah, my consent)	Jumma Prayer
3	0.58	عمران خان، وزیراعظم عمران، اعظم عمران، پاکستان عمران، خان نیازی، خان پاکستان، خان حکومت، نہیں عمران، خان اپ، خان کہا Imran Khan (IK), Prime minister (PM) Imran, great Imran, Pakistan Imran, Khan Niazi, Khan Pakistan, Khan government, No Imran, Khan you, Khan say	PM Imran Khan
4	0.38	گرل بیوی، بیوی گرل، باھر گرل، ملک ریاض، بیوی نہیں، نہیں گرل، پیش خدمت، بیوی پولیس، پاک فوج، ریاض بیٹی Colonel wife, wife colonel, outside colonel, malik Riaz, no wife, no colonel, present, wife police, Pak army, Riaz daughter	Colonel's Wife
5	0.81	علیہ وسلم، صلی علیہ، رسول صلی، اپ صلی، محمد صلی، کریم صلی، نبی کریم، حضور صلی، نبی صلی، اکرم صلی words of Darood Shareef- a specific saying to compliment Prophet Muhammad (PBUH)	Prophet Muhammad (SAW)
6	0.67	لاک ڈاؤن، ڈاؤن ختم، ڈاؤن نرمی، سمارٹ لاک، ڈاؤن نہیں، ڈاؤن وجہ، مکمل لاک، حکومت لاک، سخت لاک، اسمارٹ لاک Lockdown, down over, down flexible, smart lock, no down, reason down, complete lock, govt lock, strict lock, smart lock	Lockdown
7	0.67	نواز شریف، میاں نواز، محمد نواز، میاں محمد، مریم نواز، ایٹمی دھماکے، شکریہ نواز، یوم تکبیر، شہباز شریف، شریف پاکستان Nawaz sharif, Mian Nawaz, M.Nawaz, Mian Muhammad, Maryam Nawaz, nuclear explosion, thanks Nawaz, youm-e-takbeer, Shahbaz sharif, Sharif Pakistan	Youm-e-takbeer
8	0.69	زندہ باد، پاکستان زندہ، پاک فوج، باد زندہ، نبوت زندہ، تاجدار ختم، لبیک لبیک، فوج زندہ، باد پاکستان، پائندہ باد Zindabad, Pakistan Zindabad, Pak army, bad zinda, prophethood zinda, tajdar-e-khatm, labaik labaik, army zinda, baad Pakistan, Paindabad	TLP Treaty
9	0.58	طارق جمیل، مولانا طارق، حامد میر، جمیل معافی، معافی مانگ، جمیل میڈیا، جمیل دعا، میڈیا جھوٹا، عالم دین، اجزا چمن Tariq Jameel, Molana Tariq, Hamid Mir, Jameel Forgiveness, apologize, Jameel media, Jameel dua, liar media, Islmaic scholar, destroyed world	Molana Tariq Jameel
10	0.95	طارق عزیز، دیکھتی انکھوں، انکھوں کانوں، کانوں طارق، الیہ راجعون، انا للہ، خدا حافظ، وانا الیہ، للہ وانا، عزیز سلام Famous media host Tariq Azizs phrases of sayings and rest in peace message words	Tariq Aziz demise
11	0.35	وزیر اعظم، اعظم عمران، بلاول بھٹو، اعظم پاکستان، بھٹو زرداری، پاکستان وزیر، نہیں وزیر، اعظم ھاوس، خان وزیر، ملک وزیر PM, great Imran, Bilawal Bhutto, great Pakistan, Bhutto Zardari, Pak PM, no PM, great house, Khan PM, country PM	PM Imran Khan
12	0.59	سندھ حکومت، بلاول بھٹو، پیپلز پارٹی، وفاقی حکومت، حکومت سندھ، علی شاہ، بھٹو زرداری، مراد علی، فرقہ واریت، حکومت کارکردگی Sindh govt, Bilawal Bhutto, Peoples party, federal govt, Sindh govt. Ali shah, Bhutto Zardari, Murad Ali, sectarianism, govt performance	Sindh Govt. performance
13	0.83	عید مبارک، اھل اسلام، اسلام عید، دلی عید، اپ عید، گہرایوں عید، مبارک عید، مسلمانوں عید، عید الفطر، عالم اسلام Eid Mubarak, people of Islam, Islam eid, heartedly eid, you eid, depth eid, Mubarak eid, muslim eid, eid-ul-fitr, Islamic world	Eid-ul-fitr
14	0.52	کورونا وایرس، پاکستان کورونا، پھر کورونا، ملک پھر، وایرس مریضوں، وایرس شکار، وایرس خلاف، دنیا پھر، وایرس وجہ، وایرس متاثرہ Covid pandemic, Pakistan covid, whole covid, whole country, covid patients, virus victim, virus against, whole word, virus reason, virus affected	Covid report
15	0.35	ٹی ایی، ایی حکومت، بلاول بھٹو، بھٹو زرداری، سوشل میڈیا، پیپلز پارٹی، تحریک، انصاف، چیئرمین بلاول، ار وایی، مسلم لیگ TI, I govt, Bilawal Bhutto, Bhutto Zardari, social media, Peoples party, tehreek, justice, chairman Bilawal, RY, Muslim league	Political parties press conference
16	0.74	ختم نبوت، نبوت زندہ، عقیدہ ختم، تاجدار ختم، ناموس رسالت، نبوت ایمان، نبوت منکر، لبیک لبیک، نبوت صلی، نبوت صلی Final prophet hood, prophet hood alive, final belief, tajdar-e-khatm, namoos-e-risalat, prophet hood faith, prophet hood denial, labaik labaik, prophet hood, prophet hood	Final Prophethood
17	0.93	عطا فرمایے، فرمایے امین، اعلی مقام، مقام عطا، جنت الفردوس، الفردوس اعلی، مغفرت فرمایے، صبر جمیل، جمیل عطا، ثم امین bestow, give Aameen, highest rank, bestow rank, supreme Jannah, greatest jannah, bestow forgiveness, great patience, bestow, sum Aameen	Maghfirah prayer
18	0.74	عبدالعزیز رحمہ، علی رضی، سیدنا عمر، العزیز رحمہ، حضرت علی، حرمتی کرنے، قبر ہے، امت مسلمہ، قبر مبارک، عبدالعزیز رحمہ Abdul Aziz rehmat, Ali (RA), syed Umar, Al-Aziz rehmat, Hazrat Ali, desecrate, on grave, Muslim ummah, holy grave, abdul Aziz rehmat	Umar bin Abdul Aziz grave disrespect
19	0.41	عمران نیازی وزیراعظم عمران بلاول بھٹو نیازی چور نیازی حکومت بھٹو زرداری ٹائیگر فورس نیازی پاکستان نہیں عمران ریاست مدینہ Imran Niazi, PM Imran, Bilawal Bhutto, Niazi thief, Niazi govt, Bhutto Zardari, tiger force, Niazi Pakistan, no Imran, Madina state	Critiques on PTI Govt.
20	0.97	اہلسنت مجدد، اے امام، فاضل بریلوی، اللہ علیہ، الشاہ امام، رحمہ علیہ، بریلوی رحمہ، رضا خان، بریلوی رحمہ، اعلی حضرت Ehle-sunnah mujadid, aye Imam, Fazil brelvi, Allah eleh, Alshaa imam, Rehmat eleh, brelvi rehmat, Raza khan, brelvi remat, great hazrat	Imam Ahmed Raza Brelvi

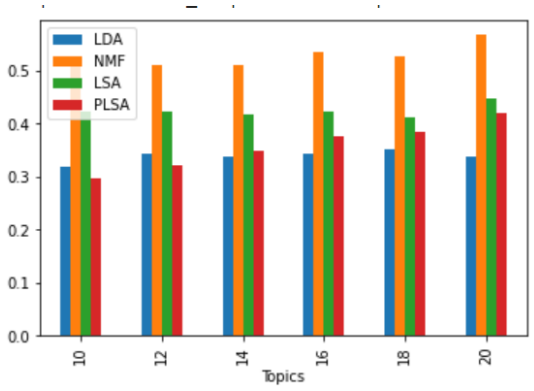

FIGURE 14. Models *C_v* coherence evaluation with bi-gram.

A. EYE BALLING MODELS

It is difficult to determine the relative size of circles and distance between circles in eyeballing models, i.e. difficulty in the distinction of topic importance than others and the extent of the difference. Secondly, information obtained from different projections is not related (e.g., t-SNE, pcoa, mmds).

B. INTRINSIC EVALUATION METRICS

Perplexity and log-likelihood do not consider semantic relations between words. On the other hand, topic coherence does that, but multiple metrics of topic coherence (*C_umass*, *C_npmi*, *C_v*, *C_uci*, etc) had different evaluation scores on the same dataset that makes them difficult to compare. Besides, a model can be worse due to extra 'junk topics' that are not accounted for in this measure.

C. EXTRINSIC EVALUATION METRICS

This method requires a benchmark dataset with true labels. Unfortunately, we are lacking benchmark datasets in the Urdu language.

D. HUMAN JUDGEMENTS

This process is really time-consuming, and conflict may occur due to differences in observations. The evaluator needs to indicate intruding words that are different from other true words of the topics. In this study, we have used both quantitative and qualitative ways to evaluate results. We have chosen *C_v* coherence, perplexity, log-likelihood, and precision (or p@n) as our intrinsic measure to evaluate discovered topics of our final model. *C_v* metric quantitatively assesses each discovered topic based on the boolean sliding window and combining the normalized point-wise mutual information and indirect cosine similarity. Also, topic coherence has been proven to match well with human judgments [38]. Precision is a commonly used metric in information retrieval [30], [55], and we have used it to evaluate the top probability words produced by the topic models in each topic with values of n = 5 and 10. We have chosen the top ten words of each topic and presented them to three human judges who

TABLE 8. LDA topics and coherence scores on dataset-H.

Topics	CV	Top 10 Words (LDA topic count=5)	Labels
1	0.59	جسٹس, فائز, قاضی, سپریم, کورٹ, ارطغرل, غازی, ریفرنس, جج, کیس Justice, Faiz, Qazi, supreme, court, Ertugurl, ghazi, reference, judge, case	Justice Qazi Faiz Isa case
2	0.76	افراد, ٹائمز, ہزار, اموات, صحت, پنجاب, کوروناوائرس, مریضوں, کیسز, مریض Persons, times, thousands, deaths, health, Punjab, corona virus, patients, cases, patient	Covid report
3	0.91	ایکسچینج, اسٹاک, حملہ, دھشت, پولیس, حملے, سلام, ناکام, شہید, جہنم Exchange, stock, attack, terrorism, police, attacks, greeting, failed, martyred, hell	Karachi stock exchange attack
4	0.93	چاند, عید, فواد, مفتی, چوھدری, منیب, سائنس, کمیٹی, چودھری, رویت Moon, eid, Fawad, mufti, chauhdhary, Muneeb, science, committee, chaudhary, ruwait	Crescent moon sighting
5	0.41	مولانا, جمیل, طارق, کشمیر,معافی, پٹرول, پیٹرول, میر, بھارت, نیازی Molana, Jameel, Tariq, Kashmir, forgiveness, petrol, Mir, India, Niazi	Unknown
		Top 10 Words (LDA topic count=10)	
1	0.92	جسٹس, فائز, قاضی, سپریم, کورٹ, ریفرنس, کنجر, ٹاک, کیس, اھلیہ Justice, Faiz, Qazi, supreme, court, reference, abusive, talk, case, wife	Justice Qazi Faiz Isa case
2	0.76	ٹائمز, افراد, ہزار, اموات, صحت, پنجاب, کیسز, مریضوں, مریض, کوروناوائرس Times, people, thousands, deaths, health, Punjab, cases, patients, patient, corona virus	Covid report
3	0.66	طیارہ, حادثے, سرور, حادثہ, پولیس, وائس, دھشت, آف, جہاز, طیارے Plane, accidents, Sarwar, accident, police, voice, terrorism, off, plane, planes	Plane Crash
4	0.93	چاند, عید, فواد, مفتی, چوھدری, منیب, سائنس, کمیٹی, چودھری, رویت Moon, eid, Fawad, mufti, chauhdhary, Muneeb, science, committee, chaudhary, ruwait	Crescent moon sighting
5	0.57	مولانا, جمیل, طارق, کشمیر, معافی, میر, حامد, سچ, بھارتی, بھارت Molana, Jameel, Tariq, Kashmir, forgiveness, Mir, Hamid, truth, Indian, India	Hamid mir remarks on M.Tariq Jameel
6	0.67	کورٹ, سپریم, چیف, جسٹس, عدالت, حکم, غیر, آف, ذمہ, شاپنگ Court, supreme, chief, justice, court, ordinance, non,off, responsibility, shopping	Supreme Court order
7	0.54	ارطغرل, طیارہ, غازی, حادثے, حادثہ, ڈرامہ, عطا, طیارے, جہاز, موت Ertugurl, plane, Ghazi, accidents, accident, play, bestow, planes, plane, death	Plane accident/Ertugurl play
8	0.54	پولیس, وزیراعلی, کوروناوائرس, عدالت, وھاب, بھٹو, درخواست, بلاول, تقسیم, صورتحال Police, chief minister, corona virus, court, Wahab, Bhutto, application, Bilawal, division, situation	Sindh Govt food distribution
9	0.61	پٹرول, پیٹرول, نیازی, مافیا, بجٹ, اسمبلی, رائے, قیمت, مہنگا, تحریک Petrol, petrol, Niazi, mafia, budget, assembly, opinion, price, expensive, movement	Petrol Price
10	0.91	ایکسچینج, اسٹاک, حملہ, دھشت, پولیس, حملے, سلام, ناکام, شہید, جہنم Exchange, stock, attack, terrorism, police, attacks, greeting, failed, martyrs, hell	Karachi Stock Exchanage attack
		Top 10 Words (LDA topic count=15)	
1	0.92	جسٹس, فائز, قاضی, سپریم, کورٹ, ریفرنس, کنجر, ٹاک, کیس, اھلیہ Justice, Faiz, Qazi, supreme, court, reference, abusive, talk, case, wife	Justice Qazi Faiz Isa case
2	0.53	سیاہ, نیکیاں, نمائندے, مشین, آو, موڑ, تصویریں, جج, استدعا, پاکستان Black, virtues, representatives, machine, come, turn, pictures, judge, appeal, Pakistan	Unknown
3	0.40	سرور, شہید, وائس, آف, غلام, تحریک, دھشت, اسمبلی, پیٹرول, رائے Sarwar, martyr, voice, off, servant, movement, terrorism, assembly, petrol, opinion	Unknown
4	0.75	کشمیر, بھارت, بھارتی, فوج, مقبوضہ, آزاد, چین, شہید, کشمیریوں, کشمیری Kashmir, India, Indian, army, occupied, free, China, martyr, Kashmiris, Kashmiri	Kashmir
5	0.61	مولانا, جمیل, طارق, معافی, میر, حامد, پٹرول, پیٹرول, سچ, مافیا Molana, Jameel, Tariq, pardon, Mir, Hamid, petrol, petrol, truth, mafia	Unknown
6	0.79	کورٹ, جسٹس, سپریم, چیف, فائز, عدالت, حکم, قاضی, جج, غیر Court, justice, supreme, chief, Faiz, court, ordinance, Qazi, judge, non	Justic Qazi Faiz Isa hearning
7	0.80	طیارہ, حادثے, حادثہ, طیارے, جہاز, عطا, موت, منزل, لواحقین, صبر Plane, accidents, accident, planes, plane, bestow, death, destination, relatives, patience	Plane accident
8	0.54	پولیس, وزیراعلی, کوروناوائرس, عدالت, بھٹو, وھاب, درخواست, بلاول, تقسیم, صورتحال Police, chief minister, corona virus, court, Wahab, Bhutto, application, Bilawal, division, situation	Sindh Govt. food distribution
9	0.52	نیازی, بجٹ, چینل, تبدیلی, پیٹرول, سرکار, دشمن, اسمبلی, سعید, چینی Niazi, budget, channel, change, petrol, govt., enemy, assembly, Saeed, sugar	Budget
10	0.53	سیاہ, نیکیاں, نمائندے, مشین, آو, موڑ, تصویریں, جج, استدعا, پاکستان Black, virtues, representatives, machine, come, turn, pictures, judge, appeal, Pakistan	Unknown
11	0.91	ایکسچینج, اسٹاک, حملہ, دھشت, پولیس, حملے, سلام, ناکام, جہنم, شہید Exchange, stock, attack, terrorism, police, attacks, greeting, failed, martyrs, hell	Karachi stock exchange attack
12	0.93	چاند, عید, فواد, مفتی, چوھدری, منیب, سائنس, کمیٹی, چودھری, رویت Moon, eid, Fawad, mufti, chauhdhary, Muneeb, science, committee, chaudhary, ruwait	Moon sighting
13	0.89	ارطغرل, غازی, ڈرامہ, ارتغرل, ڈرامے, کردار, سلطان, ترکی, ارتغل, دیکھنے Ertugurl, Ghazi, play, Ertugurl, dramas, characters, Sultan, Turkey, Ertugurl, watch	Ertugurl Ghazi play
14	0.76	افراد, ہزار, ٹائمز, اموات, صحت, پنجاب, کیسز, مریضوں, مریض, کوروناوائرس people, thousand,times, deaths, health, Punjab, cases, patients, patient, corona virus	Covid report
15	0.70	ٹائمز, افراد, ہزار, اموات, مریض, کیسز, پنجاب, بحق, متاثر, رائے Times, people, thousand, deaths, patients, cases, Punjab, die, affected,opinion	Covid report

are familiar with common knowledge about tweets, for labeling their topics. We have also asked them to mark topic non-discriminatory words that do not make any sense related to the underlying topic. A topic has been considered wrong if it contains more than half semantically incorrect words or uninterpretable by humans to relate them in

TABLE 9. LDA computed coherence, perplexity and log-likelihood on dataset-H.

BOW uni-gram				BOW bi-gram		
Topics	C_v	Perplexity	Log_likelihood	C_v	Perplexity	Log_likelihood
5	**0.743**	1452.73	-1658904.85	0.487	**883.47**	**-525226.60**
10	0.674	1416.34	-1653125.11	0.470	900.42	-526697.96
15	0.668	**1399.78**	**-1650446.07**	0.462	908.52	-527391.12
20	0.612	1441.96	-1657209.45	0.505	930.64	-529253.58
25	0.625	1465.31	-1660869.12	**0.510**	937.19	-529796.48
30	0.613	1479.65	-1663087.55	0.507	962.77	-531881.56

TABLE 10. Cohen's Kappa for judges agreement and precision words on dataset-A.

Dataset-A	Topic Labelling	Word Labelling)	
		p@5	p@10
NMF	0.80	0.95	0.93
LDA long-pseudo based	0.71	0.92	0.88

any topic. Later, we increased the number of words to 20 for taking consent from judges about decided labels of topics and computed Cohen's kappa value of these judges to analyze their agreement to ensure correct topic annotation [17]. Table 7 and 8 represent the number of incorrect topics highlighted with red color, commonly decided by judges for each approach. Table 10 represents their Cohen's Kappa score alongside average precision values. Here Cohen's Kappa score of 0.8 on topics produced by the NMF model indicates a good agreement between judges alongside the precision values on top 5 and 10 words. Whereas, 0.7 kappa score is a fair agreement between judges with their precision score.

VI. CONCLUSION AND FUTURE WORK

This paper has reported our analysis of conventional topic models on Urdu language tweets. Traditional topic models like LSA, pLSA, LDA, and HDP have performed below par, when applied directly to the short text in the Urdu language. From coherence analysis, hashtag analysis, and topic words visualizations, we selected NMF and LDA as our final models for advanced experimentation and demonstrated that generative model, like LDA gives better outcomes on the strategy of aggregated Urdu tweets by hashtag correlation analysis treated as a long pseudo document. Nevertheless, topic modeling via NMF outperformed when applied directly on Urdu tweets than other probabilistic topic models, especially with TF-IDF bi-gram vectors. In future, we intend to apply recent developments in deep learning methods after preparing embedding in the domain to infer niche topics for our Urdu language tweets.

REFERENCES

[1] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *Int. J. Adv. Comput. Sci. Appl.*, vol. 6, no. 1, pp. 147–153, 2015.
[2] D. Alvarez-Melis and M. Saveski, "Topic modeling in Twitter: Aggregating tweets by conversations," in *Proc. 10th Int. AAAI Conf. Web Social Media*, 2016, pp. 519–522.
[3] W. Anwar, I. S. Bajwa, M. A. Choudhary, and S. Ramzan, "An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution," *IEEE Access*, vol. 7, pp. 3224–3234, 2019.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[5] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," in *Proc. GSCL*, 2009, pp. 31–40.
[6] J. Boyd-Graber, Y. Hu, and D. Mimno, "Applications of topic models," *Found. Trends Inf.*, vol. 11, nos. 2–3, pp. 143–296, 2017, doi: 10.1561/1500000030.
[7] G.-B. Chen and H.-Y. Kao, "Word co-occurrence augmented topic model in short text," *Int. J. Comput. Linguistics Chin. Lang. Process.*, vol. 20, no. 2, pp. 45–64, Dec. 2015.
[8] A. Daud, W. Khan, and D. Che, "Urdu language processing: A survey," *Artif. Intell. Rev.*, vol. 47, no. 3, pp. 279–311, 2017.
[9] 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, 1990.
[10] J. Dorsey. (2020). *Twitter by the Numbers: Stats, Demographics & Fun Facts*. URL: https://www.omnicoreagency.com/twitter-statistics/
[11] L. Gui, J. Leng, G. Pergola, Y. Zhou, R. Xu, and Y. He, "Neural topic model with reinforcement learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3469–3474.
[12] P. Gupta, Y. Chaudhary, F. Buettner, and H. Schütze, "Document informed neural autoregressive topic models with distributional prior," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 6505–6512.
[13] R. He, X. Zhang, D. Jin, L. Wang, J. Dang, and X. Li, "Interaction-aware topic model for microblog conversations through network embedding and user attention," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1398–1409.
[14] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 1999, pp. 50–57.
[15] L. Jiang, H. Lu, M. Xu, and C. Wang, "Biterm pseudo document topic model for short text," in *Proc. IEEE 28th Int. Conf. Tools with Artif. Intell. (ICTAI)*, Nov. 2016, pp. 865–872.
[16] S. Kapadia. (Apr. 2019). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. [Online]. Available: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
[17] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, Mar. 1977.
[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[19] S. Lee, J. Kim, and S.-H. Myaeng, "An extension of topic models for text classification: A term weighting approach," in *Proc. Int. Conf. Big Data Smart Comput. (BIGCOMP)*, Feb. 2015, pp. 217–224.
[20] J. Li, M. Liao, W. Gao, Y. He, and K.-F. Wong, "Topic extraction from microblog posts using conversation structures," in *Proc. ACL*, vol. 1. Singapore: World Scientific, 2016, pp. 419–437.
[21] X. Li, A. Zhang, C. Li, J. Ouyang, and Y. Cai, "Exploring coherent topics by topic modeling with term weighting," *Inf. Process. Manage.*, vol. 54, no. 6, pp. 1345–1358, Nov. 2018.
[22] S. Likhitha, B. S. Harish, and H. M. K. Kumar, "A detailed survey on topic modeling for document and short text data," *Int. J. Comput. Appl.*, vol. 178, no. 39, pp. 1–9, Aug. 2019.
[23] T. Lin, W. Tian, Q. Mei, and H. Cheng, "The dual-sparse topic model: Mining focused topics and focused terms in short text," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 539–550.
[24] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, "An overview of topic modeling and its current applications in bioinformatics," *SpringerPlus*, vol. 5, no. 1, p. 1608, Dec. 2016.
[25] H.-Y. Lu, L.-Y. Xie, N. Kang, C.-J. Wang, and J.-Y. Xie, "Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1192–1198.
[26] M. Lyra. (2017). *Evaluating Topic Models, Pydataberlin-2017*. [Online]. Available: https://github.com/mattilyra/pydataberlin-2017
[27] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2013, pp. 889–892.
[28] T. Miller, D. Dligach, and G. Savova, "Unsupervised document classification with informed topic models," in *Proc. 15th Workshop Biomed. Natural Lang. Process.*, 2016, pp. 83–91.
[29] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.
[30] A. Mukherjee and B. Liu, "Aspect extraction through semi-supervised modeling," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2012, pp. 339–348.
[31] S. Munir, S. Wasi, and S. I. Jami, "A comparison of topic modelling approaches for Urdu text," *Indian J. Sci. Technol.*, vol. 12, p. 45, Dec. 2019.

[32] Z. Nasim and S. Haider, "Cluster analysis of Urdu tweets," *J. King Saud Univ.-Comput. Inf. Sci.*, Aug. 2020, doi: 10.1016/j.jksuci.2020.08.008.
[33] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Proc. Hum. Lang. Technol., Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 100–108.
[34] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Trans. Assoc. Comput. Linguistics*, vol. 3, no. 1, pp. 299–313, 2015.
[35] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Mach. Learn.*, vol. 39, nos. 2–3, pp. 103–134, 2000.
[36] C. Ordun, S. Purushotham, and E. Raff, "Exploratory analysis of covid-19 tweets using topic modeling, UMAP, and DiGraphs," 2020, *arXiv:2005.03082*. [Online]. Available: http://arxiv.org/abs/2005.03082
[37] X. Quan, C. Kit, Y. Ge, and S. J. Pan, "Short and sparse text topic modeling via self-aggregation," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 2270–2276.
[38] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, Feb. 2015, pp. 399–408.
[39] 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)*, Jan. 2018, pp. 117–123.
[40] B. Shi, W. Lam, S. Jameel, S. Schockaert, and K. P. Lai, "Jointly learning word embeddings and latent topics," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 375–384.
[41] T. Shi, K. Kang, J. Choo, and C. K. Reddy, "Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations," in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1105–1114.
[42] A. Steinskog, J. Therkelsen, and B. Gambäck, "Twitter topic modeling by tweet aggregation," in *Proc. 21st Nordic Conf. Comput. Linguistics*, 2017, pp. 77–86.
[43] Y. Sun, K. Loparo, and R. Kolacinski, "Conversational structure aware and context sensitive topic model for online discussions," in *Proc. IEEE 14th Int. Conf. Semantic Comput. (ICSC)*, Feb. 2020, pp. 85–92.
[44] G. Tao, Y. Miao, and S. Ng, "COVID-19 topic modeling and visualization," in *Proc. 24th Int. Conf. Inf. Visualisation (IV)*, Sep. 2020, pp. 711–716.
[45] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.
[46] S. Wang, Z. Chen, G. Fei, B. Liu, and S. Emery, "Targeted topic modeling for focused analysis," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1235–1244.
[47] Y. Wang, J. Liu, Y. Huang, and X. Feng, "Using hashtag graph-based topic model to connect semantically-related words without co-occurrence in microblogs," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 7, pp. 1919–1933, Jul. 2016.
[48] K. Xu, F. Liu, T. Wu, S. Bi, and G. Qi, "A fast and effective framework for lifelong topic model with self-learning knowledge," in *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data* (Lecture Notes in Computer Science), vol. 10565. Cham, Switzerland: Springer, 2017, pp. 147–158, doi: 10.1007/978-3-319-69005-6_13.
[49] G. Xun, Y. Li, W. X. Zhao, J. Gao, and A. Zhang, "A correlated topic model using word embeddings," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 4207–4213.
[50] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web (WWW)*, 2013, pp. 1445–1456.
[51] K. Yang, Y. Cai, Z. Chen, H.-F. Leung, and R. Lau, "Exploring topic discriminating power of words in latent Dirichlet allocation," in *Proc. COLING 26th Int. Conf. Comput. Linguistics: Tech. Papers*, 2016, pp. 2238–2247.
[52] J. Yin and J. Wang, "A Dirichlet multinomial mixture model-based approach for short text clustering," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 233–242.
[53] H. Zhao, L. Du, and W. Buntine, "A word embeddings informed focused topic model," in *Proc. Asian Conf. Mach. Learn.*, 2017, pp. 423–438.
[54] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic models," in *Proc. Eur. Conf. Inf. Retr.*, in Lecture Notes in Computer Science, vol 6611. Berlin, Germany: Springer, 2011, pp. 338–349, doi: 10.1007/978-3-642-20161-5_34.
[55] X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 56–65.
[56] Y. Zuo, J. Wu, H. Zhang, H. Lin, F. Wang, K. Xu, and H. Xiong, "Topic modeling of short texts: A pseudo-document view," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 2105–2114.
[57] Y. Zuo, J. Zhao, and K. Xu, "Word network topic model: A simple but general solution for short and imbalanced texts," *Knowl. Inf. Syst.*, vol. 48, no. 2, pp. 379–398, Aug. 2016.

ZOYA received the bachelor's degree in information technology from Punjab University College of Information Technology (PUCIT), Lahore, Pakistan, and the master's degree in computer science from the Government College University (GCU), Lahore. She is currently pursuing the Ph.D. degree with the National University of Sciences and Technology (NUST). Her research interests include NLP, machine learning, and deep learning.

SEEMAB LATIF (Senior Member, IEEE) received the Ph.D. degree from The University of Manchester, U.K. She is currently an Assistant Professor and a Researcher with the National University of Sciences and Technology (NUST), Pakistan. Her professional services include industry consultations, the conference chair, a technical program committee member, and a reviewer for several international journals and conferences. In the last three years, she has established research collaborations with national and international universities and institutes. She has also secured research grants from the National ICT Research and Development Grass-Root Initiative, the Higher Education Commission Technology Development Fund, and U.K. ILM Ideas. Her research interests include artificial intelligence, machine learning, data mining, and NLP. She received the School Best Teacher Award, in 2016, and the University Best Innovator Award, in 2020. She is also the Founder of NUST spin-off company, Aawaz AI Tech.

FAISAL SHAFAIT received the Ph.D. degree (Hons.) in computer engineering from the Technical University of Kaiserslautern (TUKL), Germany, in 2008. He is currently working as a Professor with the School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan. Previously, he was an Adjunct Senior Lecturer with the School of Computer Science and Software Engineering, The University of Western Australia, Perth, Australia. He was a Senior Researcher with German Research Center for Artificial Intelligence (DFKI) and an Adjunct Lecturer with TUKL. His research interests include machine learning and pattern recognition with a special emphasis on applications in document image analysis. He has coauthored over 100 publications in international peer-reviewed conferences and journals in this area.

RABIA LATIF received the bachelor's degree in computer science from COMSATS Institute of Information Technology, Islamabad, and the master's degree in information security and the Ph.D. degree in cloud-assisted wireless body area networks from the National University of Sciences and Technology (NUST), Pakistan. She is currently working as an Assistant Professor with the College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia. Her research interests include wireless body area networks, cloud computing, and information security.