IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 6, 2020, accepted May 18, 2020, date of publication May 27, 2020, date of current version June 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997907

FLAG-PDFe: Features Oriented Metadata Extraction Framework for Scientific Publications

MUHAMMAD WAQAS AHMED AND MUHAMMAD TANVIR AFZAL
Department of Computer Science, Capital University of Science and Technology, Islamabad 44000, Pakistan
Corresponding author: Muhammad Waqas Ahmed (m.waqasjanjua@gmail.com)

ABSTRACT The unprecedented growth of the research publications in diversified domains has overwhelmed the research community. It requires a cumbersome process to extract this enormous information by manually analyzing these research documents. To automatically extract content of a document in a structured way, metadata and content must be annotated. Scientific community has been focusing on automatic extraction of content by forming different heuristics and applying different machine learning techniques. One of the renowned conference organizers, ESWC organizes state-of-the-art challenge to extract metadata like authors, affiliations, countries in affiliations, supplementary material, sections, table, figures, funding agencies, and EU funded projects from PDF files of research articles. We have proposed a feature centric technique that can be used to extract logical layout structure of articles from publishers with diversified composition styles. To extract unique metadata from a research article placed in logical layout structure, we have developed a four-staged novel approach ''FLAG-PDFe''. The approach is built upon distinct and generic features based on the textual and the geometric information from the raw content of research documents. At the first stage, the distinct features are used to identify different physical layout components of an individual article. Since research journals follow their unique publishing styles and layout formats, therefore, we develop generic features to handle these diversified publishing patterns. We employ support vector classification (SVC) in the third stage to extract the logical layout structure (LLS)/ sections of an article, after performing comprehensive evaluation of generic features and machine learning models. Finally, we further apply heuristics on LLS to extract the desired metadata of an article. The outcomes of the study are obtained using the gold standard data set. The results yields 0.877 recall, precision 0.928 and 0.897 F-measure. Our approach has achieved a 16% gain on f-measure when compared to the best approach of the ESWC challenge.

INDEX TERMS Machine learning, research article, metadata extraction, text patterns, document structure analysis.

I. INTRODUCTION

Research plethora over the web increases rapidly due to millions of annual publications of research articles [1]–[3]. These cross-disciplinary publications are linked through online citation indexes so that a research community can establish the relevance to the literature. More often scholars cogitate queries based on complex scenarios to retrieve their required research documents from this colossal scientific resource. Researchers post their queries to find scholarly articles on famous online search engines like Google Scholar[1] or Semantic Scholar,[2] and renowned digital libraries like DBLP[3] or ACM.[4] However, these platforms do not hold adequate potential to intelligently process the query which results into surplus results. This is due to the fact that these search engines harness citation indexes and article's full text search to retrieve the information wherein one of the potential aspects, structural information is overlooked. Therefore, human-understandable metadata like author name, affiliation, country, email, section headings with levels, funding agency,

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

[1]https://scholar.google.com
[2]https://www.semanticscholar.org/
[3]http://dblp.uni-trier.de/
[4]https://dl.acm.org/

table, and figure caption requires indexing and storage in a machine comprehendible form to facilitate the processing of metadata-based queries. In this context, metadata extraction tools have gained popularity to extract and store machine recognizable research articles content to furnish precise semantic queries. Recently, the research community has deemed metadata extraction as a challenge. In the current era, metadata extraction from PDF files is considered as a great challenge. Every year, various efforts are put in the form of well-known conferences like SemPub,[5] CLSA,[6] OKE,[7] QALD,[8] and RecSys,[9] with an objective to improve the quality of linked data [4].

Research document structure analysis and information extraction has been a well-researched area due to increase of publications in diversified domains. Currently, information extraction methods are constructed upon machine learning and heuristic-based approaches. Machine learning techniques rely on a group of fine-tuned parameters to learn good feature representations for structure extraction. These techniques are sub-categorized into ML models built using support vector machines (SVM), conditional random fields (CRF), decision trees, and deep learning based algorithms for the feature extraction and semantic detection on text documents [5]. However, they require large tagged pre-trained dataset; it has limited aspects of natural language processing and limited performance guarantee. Initial work exhibits that heuristic-based approaches perform better because they are built on natural language processing and regular expression. These approaches are constructed on a pre-defined set of rules, and requires domain knowledge for diversified data. Therefore, the rules are required to be updated every time when documents from a new publisher are extracted.

The document layout and elements are composed on geometric location and font properties of the text, which varies for different publishers. The text in a research document has different font attributes, which can uniquely identify a group of elements. These distinct features are discussed detailed in sect 3.3. The generic features dedication contributes to develop probabilistic models in different applications, as Zare *et al.* [6] in their study investigated the influence of the features to detect community structures. We have proposed a four-staged novel approach ''FLAG-PDFe'', which uses distinct physical layout properties and generic logical layout features to transform PDF based research documents into a metadata layout aware format.

The first stage reads and extract textural information from digital-born PDF files. It reads the pdf file as raw stream of data and extract text along with text font properties encapsulated in boundary boxes that consists of geometric layout coordinates. The output is in the form of text chunks with incorrect reading order. We corrected the reading order in this stage by first identifying the column layout style of the document and then calculated the line numbers of each line by measuring the distance from neighboring text chunks. These are the physical layout properties, which are distinct in every research document. We call this the pre-processing stage that generates text block with font properties, geometric location, column styles and correct reading order. The second stage extracts the feature set which will be used by the classification algorithm to extract the logical layout structure (LLS) elements in next stage. The system processes textual and physical layout properties from extracted text content to generate generic features sets. We studied formatting styles of different publishers and proposed the set of features that can be used to extract LLS from articles of diversified layout and formatting styles. The third stage uses support vector classification [7], [8](SVC) algorithm to extract different sections of the document. For model selection, we performed systematic study on different machine learning algorithms and feature selection. The final stage performs metadata extraction from LLS/ sections identified in previous stage. This stage extracts metadata information consists of author name and affiliation, country of affiliation, supplementary material, table and figure caption, funding agency, and funded projects. This extracted metadata is stored in a csv file for comparison with start-of-the-art. We have utilized diversified and comprehensive dataset to evaluate our proposed methodology. For this purpose, ESWC-2016[10](European Semantic Web Conference) conducts a semantic challenge titled as ''Extracting information from the PDF full text of the papers'' that has provided dataset along with the gold standards available at the link,[11] which conducts semantic challenges titled as ''Extracting information from the PDF full text of the papers'', along provide publicly release benchmark datasets. We evaluated our results with the results published by the conference organizers to compare with the challenge's winner [9]. The results yields 0.877 recall, precision 0.928 and 0.897 F-measure.

The subsequent section discusses the background and demonstrates previous work in detail (sect. 2). The architecture and approaches proposed to extract metadata and section information has been comprehensively explained in the methodology section (sect. 3).

II. LITERATURE REVIEW

The metadata and structure extraction from PDF-based documents is a well-explored research area since the emergence of the initial online search engines like CiteSeerx to find scholarly articles [10]–[18]. A PDF file is stored in raw binary data form and lacks structured information tags, or metadata that identifies different layout components. It requires further processing to correct the reading order and remove intercepting objects. Another prominent obstacle is the diversified nature of the document layout styles and textural features adopted by

[5]https://github.com/ceurws/lod/wiki/SemPub2017
[6]https://2018.eswc-conferences.org/call-for-challenges/
[7]https://project-hobbit.eu/open-challenges/oke-open-challenge/
[8]https://project-hobbit.eu/challenges/om2019/
[9]http://www.recsyschallenge.com/2019/

[10] https://2016.eswc-conferences.org/
[11]https://github.com/ceurws/lod/wiki/SemPub16_Task2

different scientific publishers. Initially, document structure and content were extracted using template-based techniques but researches proposed supervised machine learning techniques and specifically linear conditional random field (CRF) [19] to replace rule-based template matching. Bijari *et al.* [20] in their study introduced a hybrid algorithm based on hueristics and clustering, using BB-BC and k-means to improve k-means shortcomings in text mining. ParsCit [21] adopted CRF to extract layout and bibliographic metadata from a research document and sectLabel further explored CRF to identify different contents of a research document. Later on, ParsCit improved its technique by adopting LSTM [22]. CERMINE [23] compared its bibliographic metadata and layout extraction approaches with popular approaches of that era and outperformed PDFExtract [24] in bibliographic information extraction. Recently, CiteCeerX[12] team introduced PDFMEF [25] that blends artifacts of their existing approaches in a framework.

A. RULE BASED TECHNIQUES

Rule-based approaches require dataset to build set of rules constructed upon natural language processing, regular expression and domain knowledge. Constantin *et al.* [26] proposed a two-stage rule-based system (PDFX) using text feature and characteristics for conversion of PDF artifact documents into XML structure. Klink and Kieninger [27] proposed a rule-based approach with combination of textual features on OCR based documents. Similarly, Déjean and Meunier [28] proposed a method for transforming PDF legacy file into a structured XML file. Ramakrishna *et al.* [29] introduced (LA-PDFText), a layout aware system to facilitate text mining in the biomedical domain. Recently, Ahmad *et al.* [9] constructed heuristics-based approach with effective combination of tagged and plain text based information extraction techniques. These approaches immensely rely on regular expressions and text pattern matching. Heuristics based approaches require predefined set of rules and text patterns to identify different elements of the research document. Hence, huge set of rules has to be maintained for diversified datasets. Therefore, the underlying problem with these approaches makes them hard to manage the overlapping rules. Furthermore, domain specified knowledge is required to apply them on a diverse dataset.

B. MACHINE-LEARNING TECHNIQUES

Supervised machine learning approaches generally use classification models where pretraining of the model is required by tagging of data based on unique features. Limited number of unsupervised machine learning algorithms are used for metadata extraction as clustering algorithms are not well suited in such cases. Granitzer *et al.* [30] investigates the use of SVM and CRF on real-world systems ParsCit and the Mendeley Desktop, for automatically extracting bibliographic metadata. Tkaczyk *et al.* [31] presented an adaptive modular workflow for extraction of metadata from born-digital scholarly articles. Huy Hoang Nhat Do *et al.* introduced Enlil [32] that uses CRF to identify authors and author affiliations and SVM to discover relationship of authors with their respective institutions. Kiss and Strunk [33] purposed an unsupervised approach to detect language-independent sentence boundaries by using abbreviations. Klampfl *et al.* [34] proposed an unsupervised approach to extract presentation optimized scientific documents without structural information. The approach extracts adjacent text blocks from the PDF file by identifying the geometrical relationship, and further classifies them to originate logical structures. Tsai *et al.* [35] used an unsupervised bootstrapping algorithm for categorization and identification of the scientific research by transforming citation contexts into coherent concepts.

Previous approaches are mostly built on the data sets of research articles that are from single publisher, hence they produce optimum test results. Their performance reduces in cases when articles from different publishers are tested. The feature sets of most techniques are not well-defined. In most scenarios, the benchmark annotated dataset is not available along with the evaluation tool, therefore, a comprehensive analysis cannot be performed. The selected datasets to evaluate our proposed technique has diversified publishing style and unique metadata requirements. FLAG-PDFe outperformed renowned techniques when evaluated on selected dataset. We have made following contributions in this regard:

1) Proposed technique generates well defined features set identified at two levels; the first is the physical layout and textual properties of individual research article, which are then used to develop the generic set of features. These features can be used to extract logical layout content of articles from publishers with diversified composition styles.
2) Our technique evaluates all the logical layout content present in an article, unlike other techniques which are task specific only.
3) It does not depend on a single feature set as in few scenarios physical properties are not extracted correctly from a PDF file.
4) We have proposed a scalable multistage framework, so the future updates can be handled at any level.
5) The technique extracts unique metadata hidden in the content of the logical layout structure.
6) The technique is evaluated using gold-standard dataset with evaluation tool, which is publicly available online.

III. METHODOLOGY

We covered hypothetical, theoretical and experimental aspects in our research methodology. Hypothetically, the content of the research documents is presented in different layout and formatting styles which makes it easier for humans to comprehend different parts and sections of a document. Most of the documents share common formatting styles which makes them easily readable.

[12]http://csxstatic.ist.psu.edu/

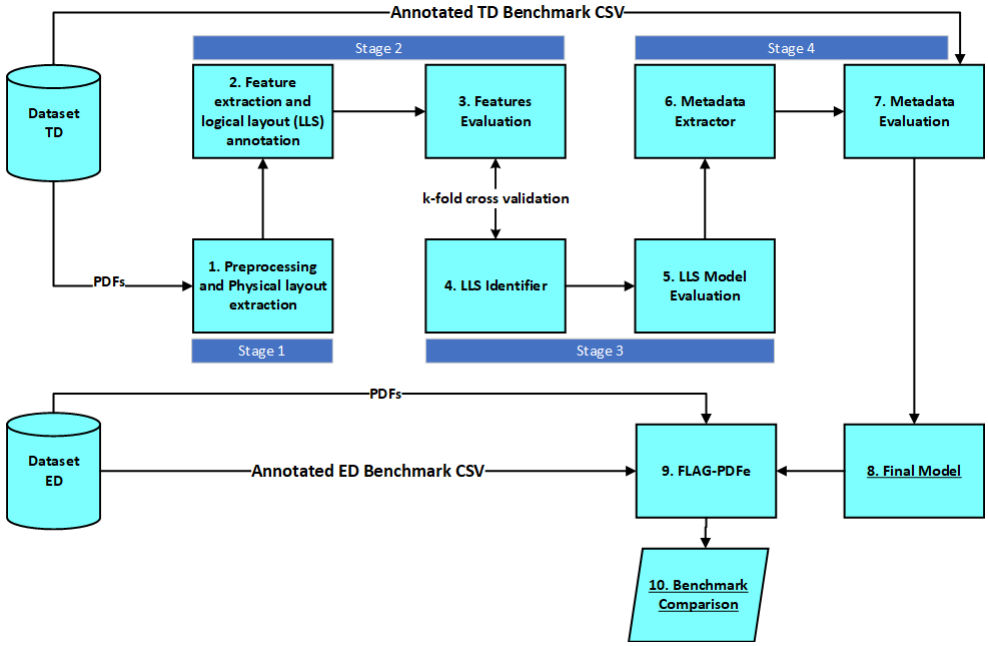

FIGURE 1. PDF base research article's metadata extraction proposed methodology flow diagram.

Theoretically, we have analyzed different formatting styles of publishers and established that these layout and formatting styles can be used to extract metadata from research articles. Since this important information and layout components require annotations, therefore, we have categorized the formatting styles into two types of structural components, one is the physical layout and other is logical layout structure components. The physical layout is based on individual article's distinct features, which consists of textual properties, geometric boundaries, paragraphs, column styles, floating object, headers and footers etc. The logical layout structures (LLS) are generic formatting features to identify different parts, contents and sections of an article that are required by the publisher. The system and proposed methodology flow diagram is shown in Figure 1.

FLAG-PDFe takes research article as an input in PDF format. The first stage extracts physical layout of a PDF file and text chunks along with geometrical location and font property. These text chucks are processed and organized in the form of text blocks, with correct reading order and document formatting style aware. In the second stage, text blocks geometric and textual properties are used to create feature sets, which are used for classification algorithm to extract logical layout structure (LLS) components of the research articles in the third stage. Finally, heuristics are applied on LLS to get desired metadata, which is sorted and stored in csv output form. In preceding sections of the paper every process is explained in their chronological order, and in the next section, the formulation and extraction of the textural information is discussed.

A. DATASET

In order to develop a comprehensive model which can be used on diversified publishing styles, we chose ESWC 2016 challenge task 2 published dataset. Various gold standard datasets from ESWC challenge are available at the link[13] along with an evaluation tool. This dataset consists of research articles having diversified format and styles adopted from publishers like ACM, LNCS, and IEEE. The dataset has two parts, first is the training dataset (TD), which consists of 45 research articles and second part test dataset (ED) consists of 40 research articles. Initially, we used training dataset (TD) of ESWC for model construction. The evaluation of model was done on test dataset (ED). The output of the ED contains of 320 CVS formatted files. We evaluated the output of the proposed system at different stages on bases of comparison done with gold standard dataset. ESWC 2017 challenge task 2[14] published a test dataset (TD) containing 40 research articles. Conference organizers have not published evaluation dataset (ED) along with evaluation of proposed techniques. However, we have also used TD dataset to further evaluate the performance of our proposed model.

B. PHYSICAL LAYOUT AND PRE-PROCESSING

1) PHYSICAL LAYOUT EXTRACTION

A PDF file is composed of raw binary data without any associated metadata and logical structural information that identifies different layout categories of the content.

[13]https://github.com/angelobo/SemPubEvaluator
[14]https://github.com/ceurws/lod/wiki/SemPub17_Task2

Therefore, the first process is to extract the textual information from the PDF file. At this stage, we used itext [36] open source java library that provides faster and reliable method to extract PDF file. Unlike other processing tools that extract text as text glyph or stream of characters, itext extracts chunk of textual elements that reduces resources and computational cost. Further, itext implements advanced strategy to extract structural components that are text chunks, font properties, geometric locations, raster images, page numbers, and vector graphics. The text chunks are retrieved, encapsulated in boundary boxes that identifies their geometric position in form of (x, y) coordinates on the page along with height and width. The itext library returns font attributes like font name, font size, bold, italic, orientation etc. We used these attributes to generate font properties feature set.

2) COLUMN STYLE IDENTIFICATION

The research documents are composed in single or double column style. This process identifies the column style of the document in order to determine the boundary of the main text body. The column layout style further helps to identify the geometric position and layout properties of the text blocks. The process first calculates the right and left outermost margins of the page. The left outermost margin is calculated by the MODE of minimum values of text blocks geometric start point, and the right outermost margin is calculated by the MODE of maximum values of text blocks geometric end value. Thereafter, the process calculates the number of columns present in the document. The process starts from left outermost margin and calculates MODE of maximum values of text blocks geometric end value. If the value is equal to right out most margin, the process stops, else process again computes the MODE of minimum values of text blocks geometric start point, till it finally reaches to the left outer margin. This process also helps to identify text blocks present in the form of decorations and footnotes.

3) CORRECT READING ORDER

Earlier systems use heuristics-based X-Y cut algorithm and KNN based Docstrum algorithm to correct the reading order of the document. Although, the output of the itext library is mostly in correct rendering order but it contains irregularities while extracting the text reading order. The reading order irregularity is due to in-text citations, algorithms, tables content, vector graph-based figures content, special characters and floating text objects. We corrected the reading order of text chucks using neighboring text geometric distance and rendering order. The process derives words from received scattered chunks of text and on the basis of geometrical location and physical distance among them. The words grouped together to formulate lines while retaining the text features of individual text chunk. This process produces plain texts having no relationship between words and lines and paragraphs. The line numbers are assigned by computing the reading order and rendering order of the content of text blocks, text with same geometrical position and column had same line numbers.

C. FEATURE EXTRACTION

The logical layout consists of font and formatting style. We have developed the approach for our system that can recognize different components of the document, based on font textual and geometric properties. We have analyzed all the possible layout variants present in the training data-set and, built the features set based on those textual properties. The features set is used by machine learning model in the next stage.

1) PHYSICAL AND LOGICAL LAYOUT EVALUATION

The logical layout determines the document's layout components comprised of title, authors and affiliation, figures, tables captions, heading and levels, paragraphs, bibliography etc. A PDF file most often lacks metadata tag associated to an individual logical layout category to support automatic retrieval or identification of required content. We have developed a framework to address this core issue by extracting the logical structure categories of PDF-based research articles to generate layout aware output.

a: FONT PROPERTIES

The itext library extracts the font properties of the text characters or chunks from a pdf file. However, the font name contains all the font information having concatenated itext font code, font family name, bold or italic information, which requires further processing to extract individual font properties. Most often, individual categories of text blocks possess different font features, like section headings composed in bold or italic to mark as prominent. Based on font families there is variation in identification of the bold and italic properties. As "Times", "Arial", "SegoeUI" and "Nimbus" etc. font families contain "bold" or "italic" keywords and Computer Modern "CM" fonts has "BX" for bold and "TI" for italic font style as represented in Figure 2.

Output Font Name of itext library	Document font Style
DRPUQI+CMBX12	Introduction
SJXVJJ+Times-Bold	ABSTRACT
PIGZEP+CMBXTI10	*Relation Intersection.*
Times New Roman,BoldItalic	***Abstract***

FIGURE 2. Font names and style extracted by itext library for document font style.

b: NEIGHBOR DISTANCE

The reading orders helps to assign line numbers to individual text line. The line number enables the system to identify the sorted order of main body content, sections heading and bibliography. However, the sorting of table numbers and

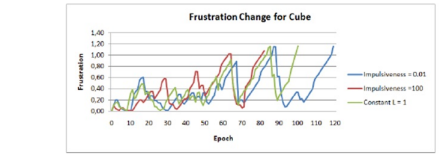

Fig. 4. Frustration Rate Changes For Block Grasping with Methods M_3 and M_4.

lation generated 116 possible grasp pair candidates that were subsequently used by the robotic arm. Since the pin is quite light, the arm pulls it down for some grasp pairs. When the pin fells down, the frustration threshold is decreased for

(a)

(Z5) Time	(–)	(–)	(–)	(–)	–	–
(Z6) Location	(–)	(+)	+	(–)	+	+

Table 1. Evaluation of the related models w.r.t. to their modeling capabilities. (- not considered, (-) partially considered, (+) implicitly provided, + fully provided)

explain how those locations can be related to each other, to create hierarchical location models.

(b)

FIGURE 3. The text blocks share common text properties, but the table caption distance from paragraph is more as compared to distance among paragraph lines.

figure numbers cannot be guaranteed. The feature measuring distance of line from top and bottom of page calculates the sorting order of the content; also, this feature enables the system to identify text blocks that are composed close to the far boundaries of the page. The distance from adjacent lines helps to identify continuity among text block to form paragraphs. However, this parameter is conclusive when the Font properties are same between distinct text blocks. Like section heading, figure and table caption have same text size and font properties as compared to body text as illustrated in Figure 3.

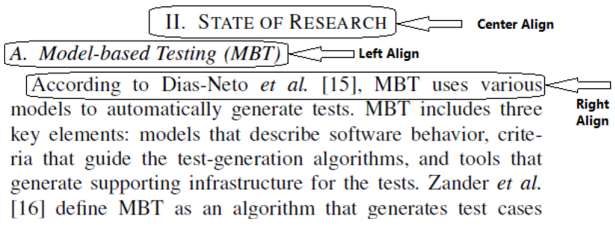

FIGURE 4. Alignment feature of the text blocks within a column.

c: TEXT LOCATION

The column style helps to determine the text location features of text blocks based on the presence of text block or line in a column. During the identification process of external boundaries, the single or double columns styles were identified. In this stage, the text blocks location in a column is defined. The documents with single column style have text blocks existing in column number one. However, with double column style a text block can exist in column number one or two, and text block that does not reside in any column is assigned with column number zero like title etc. The In Column feature has the information regarding column number of a text block. The align feature identifies the left, right or center alignment of text block with in a column. Figure 4 represents the identification of alignment of text blocks where the main section heading is center aligned within a column and starting line of each paragraph is right align and rest lines are left align. The distance of starting point of text line with reference to column start is present in start indent feature and ending point distance from column end is present in end indent feature.

d: FONT TYPOGRAPHY

The font typographical features facilitates in the identification of title, section headings and levels. Research articles have section headings in different typographic where heading has text in capital case or title case format. Here, the identification of initial capital words require some pre-processing as they may contain prepositions in small case letters. Therefore, we excluded the prepositions and then checked the initial capital phrases. In Figure 5, these text case features are found in the section heading or title of research article. Another, important typographic feature is the initial numeric values to define the heading number or heading level. The heading numbers are defined either by a numeric value or a roman value as shown in Figure 5. The sub headings in such scenarios have outline numbering styles, the system counts the number of dots and eliminates if it is present at the end of the number hierarchy.

Features	Font Typographic Styles
Main-heading with Numeric Title Case	2 Related Work
Sub-heading with Numeric Title Case	2.1 NER on Social Media for English
Main-heading with Roman Capital Case	I. INTRODUCTION
Main-heading with Numeric Capital Case	1 INTRODUCTION
Sub-heading with Numeric Title Case	2.1.2 *Dataset Limitations*

FIGURE 5. The typographical styles used in articles to indicate section headings with levels.

e: LEXICAL PROPERTIES

The research documents have meaningful content that enables the system to identify the logical layout components. It has been observed that keywords-based search like "Abstract", "Reference", "Bibliography", "General Terms", "Keywords" and "Acknowledgments" etc. can be an effective method to identify the relevant sections. Therefore, the content before "Abstract" most often contains the

title section of the document. Similarly, the content after the ''reference'' heading will have bibliographic information. The Acknowledgment section contains the funding sponsors for the project, and we shall use it in later part to extract funding Agency. The figure or table caption always starts with keywords like ''Fig'', ''Figure'', ''Viz'', ''Graph'', ''Tab'' and ''Table'' etc. However, such keywords may exist at start of a paragraph but combination of these keywords along other textual feature can be helpful in identifying captions. Email's always have ''@'' character and efficiently build regular expression on text lines with these special characters can detect correct emails addresses.

D. LOGICAL LAYOUT STRUCTURE EXTRACTION

Logical layout structure (LLS) defines the layout of an article content and all research publishers provide guideline to authors to follow their layout and formatting style. The LLS components mostly includes Title and authors section, Section headings (TOC), headers and footers, table and figure captions, and reference section. Different publishers adopt diversity formatting styles to mark these LLS components. Therefore, we have proposed a generic set of features that can identify these variations, so that machine learning algorithms can effectively extract LLS for different publishers. In this section, LLS components extracted by our proposed approach are presented.

The authors, affiliation and country of affiliation are present in authors section. After manual evaluation of the research documents we identified that authors section is located after the article title and before abstract section on the first page of the document. Therefore, first part of structure extraction is the identification and labeling of Title, authors, author's affiliation, email and country of affiliation. This section contains salient font, geometric and lexical based characteristics, which are helpful in its content identification.To develop the model, the features were assigned to these text blocks and sections were labelled.

The table of contents (TOC) and textual paragraphs are the major components of an article. The system needs to identify the headings of each section. The heading font and geometric features are different from body font features. Therefore, font features facilitate in the identification of heading text. These features are comprise of capital/ bold and italic font, geometric distance from previous and current section body, geometric distance from column's border, or numeric or roman initials.

In research articles, mostly captions are present above or below the main body of tables and figures. The captions explain the content of their associated tables and figures. The captions do have dissimilarities from main body text. However, some styles contain caption text properties similar to body text and caption number has dissimilar font properties. This posted a challenge for annotation based on overlapping features, therefore we only tagged the initial value of the figure or table caption before sequence number. We additionally used keyword phrase feature to discover table and figure captions. Keyword phrase marked by matching initial word of text chunk with any of the matching keywords like Table, Tab, Figure, Fig and Viz. We have used combination of these approaches for efficient extraction of the Table and Figure captions.

The acknowledgments section mostly has a heading ''Acknowledgment'' and has heading textual style. This section recognizes the funding agencies and individual helped materialize the research work. We used both textual style and keywords to identify this section's heading and further complete acknowledgement body is marked for next stage to extract funding agency and EU project information.

1) DISCUSSION

Before the setup of a machine learning algorithm to extract logical layout structural components from diversified layout styles. Theoretically, a few points are considered regarding the properties of the features and the nature of the problem. As earlier described the features are of different datatypes like numerical, nominal and boolean. The problem and properties shows a non linear relationship between the features. Its a classification problem with multiple class labels. The number of features are lesser then the training data instances $n \gg p$. Based on the facts described, we only selected the machine learning algorithms for evaluation to prove our theoretical evidence that best fits for non linear, distinct features and multi class labels, on a large dataset. In succeeding subsection, we shall present a brief overview of different machine learning algorithms that we evaluated for our proposed methodology, as comprehensive details and computational complexities are available [37]–[41].

2) THEORETICAL EVALUATION

The evaluation is based on n, representing the number of the training samples, where p is number of the features. For tree base classification algorithms n_{trees} represents the number of the trees. Similarly, the number of the support vectors is donated by n_{sv} and finally, n_{li} is the number of neurons at a layer i in a neural network.

Naïve Bayes algorithm depends on the conditional probability based on Bayes theorem, and generates a tree based on probability known as Bayesian Network. It's characteristics are independent from each other. The time complexity is linear for both testing $\mathcal{O}(n * p)$ and prediction $\mathcal{O}(p)$ of the model. Posterior probability is calculated by $P(A|B)$ where,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

k-nearest-neighbor are defined in the terms of distance of all instances that correspond to the point in n-D space. It searches the pattern space of close unknown tuple for k training and classify it by a majority vote of its neighbors. The distance metrics, such as Euclidean distance, Manhattan and Minkowski are used to define ''closeness''. The time complexity can be reduced to a constant $\mathcal{O}(1)$, independent of training dataset of $|D|$.

The decision tree classifier, constructs the tree based on entropy and information gain by using ID3 algorithm unlike standard deviation reduction method. The nodes represents the attributes needed to be classified, while the branches represent the allowed value. A full homogeneous sample achieves entropy equal to zero by diving the sample into equal parts. The time complexity to train the classifier is $\mathcal{O}(p * nlogn)$, and for prediction is $\mathcal{O}(p)$.

The ensemble classifiers use combination of models to increase the accuracy [42]. Different methods can be applied, where improved model M^* is created with combine series of k learned models $\{M_1, M_2, M_3, \ldots .M_k\}$ on data D with k learned sets, $\{D_1, D_2, D_3, \ldots .D_k\}$. The bagging method considers majority vote by models to improve the accuracy and the term bagging origins from ''bootstrap aggregation''. In Adaboost (Adaptive boosting), assigns weights to each classifier's vote for each training tuple to boost the accuracy of learned method. The weight is calculated on errors due to misclassification and subsequent model focus on classified tuples. Weight is calculated using $\log \frac{1-error(M_i)}{error(M_i)}$. Stacking is a heterogeneous ensemble that consists of different models. The idea is to combine predictions of the base learners (level-0), do not just vote and provide as an input to meta learner level-1 models. The Random forest ensemble the decision tree classifiers so that the collection of classifiers is a ''forest''. Each tree depends on the independently sampled values and all the trees has same distribution in the forest. The accuracy is achieve using each tree's vote and the most popular class is returned. The time complexity of bagging is $\mathcal{O}(T * t)$, where T is number of iterations and t is the average time complexity of each model. The time complexity of Adaboost is $\mathcal{O}(T * f)$, where f is the complexity of the weak learner. The time complexity of stacking is $\mathcal{O}(E_1 + E_2 + E_3 + \ldots + E_k)$. And the time complexity of random forest is $\mathcal{O}(ntree * p * d * n)$, where d is the depth of the tree.

The support vector machines classifies both linear and nonlinear data. It transforms the training data into a new higher dimension by using nonlinear mapping and searches for linear hyperplan. SVM uses support vectors to find this hyperplan. The tuples of different classes are separated using ''decision boundary'' or margins. The maximum distance between margins and classes are drawn. Finding maximum marginal hyperplane (MMH) and support vectors makes it a quadratic optimization problem. For linear data, linear SVM is employed and for nonlinear data SVM provides a bag of $K(x, x')$ kernel tricks.

$$
\begin{aligned}
&linear : \langle x, x' \rangle \\
&polynomial : (\gamma \langle x, x' \rangle + r)^d \\
&Gaussian/RBF : \exp(-\gamma \parallel x - x' \parallel^2) \\
&sigmoid : \tanh(\gamma \langle x, x' \rangle + r)
\end{aligned}
$$

where d is specified by keyword degree, r by coef0 and γ is specified by keyword gamma, must be greater than 0. The overall training time complexity for kernel method is $\mathcal{O}(n^2 p)$, and prediction time complexity is $\mathcal{O}(n_{sv} p)$ [43].

The Neural networks [44] are non deterministic algorithms that generalizes well but have minimum mathematical foundation. They are learned in an incremental fashion, and nontrivial multilayer perceptrons are used to perform complex functions. Supervised, unsupervised and reinforcement are three main types of artificial neural networks. The time complexity is $\mathcal{O}(n * e(\sum_{i=1}^{h-1} nl_i nl_{i+1}))$, where e is epochs and h is total number of layers in a neural network.

E. METADATA EXTRACTION

This is the final stage to identify metadata and structural information of the research document. This desired metadata is extracted from different logical layout structural (LLS)components. This section applies heuristics on the content of LLS to extract metadata and stores them in machine comprehensible form in order to perform task specialized queries.

1) AUTHOR AND AUTHOR AFFILIATION

In previous section, we have used machine learning approach to identify different elements of authors section. It has been observed that this information is available in three style formats.

1) ''Sequence of author names separated by commas' or tab spacing, then sequence of affiliations''.
2) ''Sequence of author names with numeric or symbols separated by commas' or tab spacing. Then sequence of related affiliations with numeric or symbol''.
3) ''Group of an author's name, author's affiliation, and email address''.

The itext library provided an edge here, as the output text rendering is in the sequence of above mentioned format styles. We applied parser based on regular expression in order to separate authors and assigned them reference ID. This id is based on sequence of rendering, numeric and symbols. Thereafter, the affiliations are assigned with authors id's based on sequence of rendering, numeric and symbols. The process generated a bipartite graph of authors and affiliations.

2) COUNTRY OF AFFILIATION

A knowledge-based library is employed having country names, city names and country domain name like de, uk etc. After retrieval of author's affiliation, the country name and city names are extracted based on comparison made with knowledge-based library. If that affiliation has missing country information, then we parse email id domain name and compared it with country domain name to extract county of affiliation. Finally, distant list of countries is stored.

3) HEADING LEVEL 1

Heading levels identification is a challenging task, different complex models proposed in the literature, efficiently identifies the table of contents. To extract table of content of a book, heuristics based on TOC identification methods using information present in TOC section are employed. However, such

TABLE 1. The performance matrix of models to extract each LLS component on training dataset.

	Naïve Bayes	SVC	KNN	Decision Tree	Extra Tree	Bagging	Adaboost	Stacking	Random Forest	Rule (stratified)
Author	0.679	0.980	0.794	0.905	0.905	0.915	0.349	0.215	0.930	0.145
Acknowledge	0.000	1.000	1.000	1.000	0.176	1.000	0.000	0.000	1.000	0.120
Authors Affiliation	0.781	0.884	0.795	0.885	0.944	0.925	0.639	0.600	0.935	0.159
Email	0.524	0.897	0.650	0.748	0.768	0.835	0.000	0.036	0.814	0.069
Figure	0.128	0.955	0.696	0.801	0.797	0.829	0.518	0.741	0.844	0.069
Heading	0.449	0.945	0.883	0.945	0.949	0.960	0.689	0.820	0.954	0.475
Table	0.156	0.934	0.294	0.578	0.533	0.643	0.000	0.000	0.552	0.030
Avg F-Measure	**0.388**	**0.942**	**0.730**	**0.837**	**0.724**	**0.872**	**0.314**	**0.345**	**0.861**	**0.152**
Avg Recall	**0.429**	**0.921**	**0.723**	**0.839**	**0.707**	**0.874**	**0.353**	**0.351**	**0.844**	**0.153**
Avg Precision	**0.517**	**0.966**	**0.747**	**0.840**	**0.759**	**0.871**	**0.316**	**0.387**	**0.887**	**0.153**

approaches are not suitable for research articles. In the previous section, the level 1 headings were annotated along with level 2 and level 3 heading and output was based on classification model. However, ESWC challenge task is only to identify level1 heading, therefore no further processing is done on output of previous stage and extracted heading is stored in ascending order.

4) TABLE AND FIGURE CAPTION

In the previous section, classification model identifies the start point of table or figure caption before the sequence number. At this stage, the remaining text chuck is analyzed. The process starts from the sequence number of table or figure and breaks when next text line has different line spacing, by which multiple lines and different text properties do not break the complete caption sentence. The system further stores the caption of table and figure in ascending order based on the sequence number.

5) SUPPLEMENTARY MATERIAL

The identification of supplementary material is part of ESWC challenge and this information is present in the footnotes. The textual properties are different from main text body properties and starts with a numeric or footnotes symbol identifier. The supplementary material is in the form of URLs. We have converted all text of footnotes in a single text block and then utilized a URL parser [15] using regular expression which extracts the complete URL from descriptive part.

6) FUNDING AGENCY AND FUNDED PROJECTS

The acknowledgments section contains the funding agencies and funded project information. We have used task specific knowledge-based approach to identify funding agency name and funded project name. The training dataset TD is analyzed and a regular expression is developed to extract funding agency by locating keywords starting with 'contri', 'support', 'fund', 'grant' and ends with 'from' or 'by' and the expression ends with ''brackets'', ''quotation marks'' or ''punctuation marks''. The Parser recognizes the funding agency name along with its acronyms and finally removes the preposition and punctuation around the funding agency information.

The final metadata extracted by our system is the list of funded projects. After manual analysis of the content we observed that this information is also available in the acknowledgement section and placed after the funding

[15] https://docs.python.org/2/library/re.html

TABLE 2. Features associated to text blocks in order to identify logical layout content of research paper.

Feature Name	Description
Font Properties	
Font Name	Name of the font (e:g Times New Roman)
Font Size	Floor rounded size of font.(e:g 11)
Is Bold	The font style is bold
Is Italic	The font style is italic
Font Orientation	Horizontal or vertical orientation of font
Text Location	
Align	Text block aligned from column (left, center, justify or right)
Column Number	Text block exists inside column boundaries
Start indent	Intending space from start of column
End indent	End of line distance from end of column
Neighbor Distance	
Line Number	The line number of the text
Page Top Distance	The line distance from the top of page
Page Bottom Distance	The line distance from the bottom of the page
Previous Line Distance	The geometric distance from the previous line
Next Line Distance	The geometric distance from next line
Font Typography	
All Capital	The text has all capital letters
Initial Capital	The text has initial capital letters
Initial Numeric	The text start with numeric
Initial Roman	The text start with a roman value
Numeric dots	The number of dots in numeric values (e:g 1.1.1 = 2)
Lexical Features	
Special Characters	Contains special characters like (:, @ ; [ ] - etc)
Is Figure	Starts with keywords (Fig, Figure, viz etc)
Is Table	Starts with keywords (Tab, Table etc)
Keyword	keywords (Abstract, Reference, Bibliography etc)
Last Character	The last character of the line.

TABLE 3. The final results and the Confusion matrix of extracted metadata by FLAG-PDFe using evaluation dataset.

	Author	Affiliation	Country	Supp-Material	Sections	Table	Figure	Funding	Projects
Authors	107	8	1	0	0	0	0	0	0
Affiliation	9	44	2	0	0	0	0	0	0
Country	0	2	41	0	0	0	0	0	0
Supp- material	0	0	0	12	0	0	0	0	0
Sections	0	0	0	0	269	0	0	0	0
Table	0	0	0	0	0	33	0	0	0
Figure	0	0	0	0	0	0	103	0	0
Funding	0	0	0	0	0	0	0	20	1
Projects	0	0	0	0	0	0	0	2	5
Actual	118	50	46	14	275	37	110	24	7
Recall	**0.907**	**0.880**	**0.891**	**0.857**	**0.978**	**0.892**	**0.936**	**0.833**	**0.714**
Precision	**0.930**	**0.800**	**0.953**	**1.000**	**1.000**	**1.000**	**1.000**	**0.952**	**0.714**
F-Measure	**0.918**	**0.838**	**0.921**	**0.923**	**0.989**	**0.943**	**0.967**	**0.889**	**0.714**

Avg Recall	Avg Precision	Avg F-Measure
0.877	**0.928**	**0.897**

agency name if available. The funding project name is placed between or after the keywords ''the'' and ''project'' like ''by the EU FP7-ICT-2011-8 project''. The regular expression finally removes the keywords and parses around the content.

IV. RESULTS

To evaluate the results, standard evaluation measures like recall, precision, and f-measure are mostly employed. These methods are based on classification parameters known as *true positive* (TP), *false positive* (FP), *true negative* (TN), or *false negative* (FN). Recall (sensitivity) is a statistical measure used to judge the relevant results produced by the model. Precision analyzes the quality of results. F-measure is the harmonic mean to measure test quality based on Recall and precision.

$$Recall\ \ \rho_i = \frac{TP_i}{TP_i + FN_i},$$
$$Precision\ \ \pi_i = \frac{TP_i}{TP_i + FP_i},$$
$$F - Measure\ \ F_i = \frac{(2 * \rho_i * \pi_i)}{(\pi_i + \rho_i)}$$

A. PHYSICAL LAYOUT STRUCTURE

We have evaluated different classification-based machine learning algorithms on the given training dataset (TD). We have performed comprehensive evaluation of each machine learning approach using confusion matrix parameters.

1) EXPERIMENTAL SETUP

The training of the models is performed by using k-fold cross validation technique, where k = 10 produced optimum result. In order to improve the performance and efficiency of the model, we have performed feature reduction by first converting categorical values to numeric values while excluding non- convertible values, and used chi-squared (chi^2) to select K best features. We trained and tested the selected models described in subsection of theoretical evaluation III-D2. The euclidean distance method performed better to find k-NN, where $k = 5$ produced optimum results. We further evaluated ensemble classifiers bagging, Adaboost and Stacking with input of classification models used in current experiments.

We have followed the guide lines of [45] for the construction of the SVM model. We set different kernel functions like linear, polynomial, Gaussian-RBF and sigmoid. In order to avoid the issue of over fitting, we choose the C value of 1 and γ value equal to 10, and by selecting Gaussian-RBF as kernel function, produced the optimal results among all the classification algorithms that we evaluated for our approach.

The Table 1 illustrates the comparison of average recall, precision and f-measure of all the classification models using TD. The results show the support vector classification (SVC) classified correctly more relevant structural components. The Table 2 shows the details of features that are finally selected for extraction of LLS components by our selected classification model. The output of this stage will be used to evaluate content present in related sections and final metadata will be generated. It also reveals that our generic feature set extraction approach has played a pivotal role to correctly identify the logical layout structure ''on the fly''.

B. METADATA INFORMATION EXTRACTION

In this section, results of extracted metadata in the document are presented. We have evaluated authors and author's affiliation, country of affiliation, sections (heading level 1), Table and Figure Captions, supplementary material, funding agency and funded project. The recall, precision and f-measure are measure of each element and the mean value of these measuring methods are calculated against each metadata element. The final model results presented in Table 3 reveals that average recall = 0.877, precision = 0.928 and F-Measure = 0.897.

Finally, we have compared our results with start-of-the-art on gold standard [46]. The previous approaches and FLAG-PDFe used same dataset and are evaluated for same evaluation parameters. In Figure 6, our approach is compared with state-of-the-art, and results suggested that our approach showed significant improvement from previous approaches, and the results indicate that FLAG-PDFe has 16% performance gain on the SemPub2016 winner.

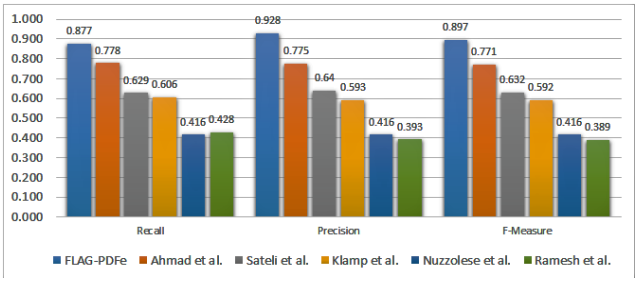

FIGURE 6. The final result comparison of FLAG-PDFe with the SumPub2016 challenge participants.

TABLE 4. The Performance Matrix of FLAG-PDFe on the TD of SemPub2017 challenge.

Metadata	Recall	Precision	F-Measure
Author	0.932	0.932	0.932
Affiliation	0.761	0.823	0.791
Country	0.837	0.953	0.891
Supp. Material	0.715	1.000	0.833
Sections	0.953	1.000	0.976
Table caption	0.923	0.923	0.923
Figure Caption	0.899	0.961	0.929
EU Projects	0.643	1.000	0.783
Average	**0.833**	**0.949**	**0.860**

Additionally, we evaluated the performance of our proposed framework on the TD consisting of 40 research paper from SemPub2017 challenge. On the bases of our TD dataset from SemPub2016, we evaluated results on 7 parameters that our technique extracts. The results presented in Table 4 reveals consistent performance of model that average recall = 0.833, precision = 0.949 and F-Measure = 0.860.

V. CONCLUSION

In this paper, we have proposed a comprehensive framework "FLAG-PDFe" for the extraction of metadata from PDF based research documents. The system converts the PDF file into metadata annotated files using classification model and heuristics. The system extracts text blocks, typography, and geometric information from a PDF raw file and reshape these features to identify and extract the logical layout structure and metadata of an article. The proposed approach consists of a novel four-stage process. The first step, the distinct features present in an individual document, are identified to extract physical layout of an article like main text boundaries, column style, and reading order etc. and further pre-processing is done to segregate paragraphs and floating objects. The second stage develops the generic features using physical layout, typographic and geometric information, which can be mapped on diversified publishing styles. In the third stage, we evaluated different machine learning methods and generic features to extract logical layout structure (LLS), the experiments reveal that support vector classification (SVC) algorithm performed best with the proposed generic set of features. Finally, the logical layout structure is further analyzed to extract desired metadata based on knowledge based and heuristics. The system outperformed previous approaches, when evaluated on gold standard (CEUR dataset).

Our study established the fact that each research article has its distinct physical layout properties, although it follows the formatting guidelines of the publishing conference or journal. Publishers use layout and formatting styles to differentiate different logical layout structural (LLS) components, therefore, generic set of features can be developed to identify logical layout components or sections for diversified publishers. The proposed approach develops both distinct and generic features used by classification algorithm on the fly, in order to recognize varying publishing styles. In future, we intend to extend stage four by extracting metadata information like subsections, bibliography, and publishing information by employing novel algorithms for natural language processing (NLP) and evaluation on additional editors.

REFERENCES

[1] A. E. Jinha, "Article 50 million: An estimate of the number of scholarly articles in existence," *Learned Publishing*, vol. 23, no. 3, pp. 258–263, Jul. 2010.
[2] L. Bornmann and R. Mutz, "Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 11, pp. 2215–2222, Nov. 2015.
[3] F. Ronzano and H. Saggion, "Knowledge extraction and modeling from scientific publications," in *Proc. Int. Workshop Semantic, Analytics, Vis.* Cham, Switzerland: Springer, 2016, pp. 11–25.
[4] A. Dimou, S. Vahdati, A. Di Iorio, C. Lange, R. Verborgh, and E. Mannens, "Challenges as enablers for high quality linked data: Insights from the semantic publishing challenge," *PeerJ Comput. Sci.*, vol. 3, p. e105, Jan. 2017.
[5] K. Bijari, H. Zare, E. Kebriaei, and H. Veisi, "Leveraging deep graph-based text representation for sentiment polarity applications," *Expert Syst. Appl.*, vol. 144, Apr. 2020, Art. no. 113090.
[6] H. Zare, M. Hajiabadi, and M. Jalili, "Detection of community structures in networks with nodal features based on generative probabilistic approach," *IEEE Trans. Knowl. Data Eng.*, early access, Dec. 17, 2019, doi: 10.1109/TKDE.2019.2960222.
[7] S. S. Keerthi and C.-J. Lin, "Asymptotic behaviors of support vector machines with Gaussian kernel," *Neural Comput.*, vol. 15, no. 7, pp. 1667–1689, Jul. 2003.
[8] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, Mar. 2002.
[9] R. Ahmad, M. T. Afzal, and M. A. Qadir, "Information extraction from PDF sources based on rule-based system using integrated formats," in *Semantic Web Evaluation Challenge*. Cham, Switzerland: Springer, 2016, pp. 293–308.
[10] J. Wu, K. M. Williams, H.-H. Chen, M. Khabsa, C. Caragea, S. Tuarob, A. G. Ororbia, D. Jordan, P. Mitra, and C. Lee Giles, "Citeseerx: Ai in a digital library search engine," *AI Mag.*, vol. 36, no. 3, pp. 35–48, 2015.
[11] M. Granitzer, M. Hristakeva, R. Knight, K. Jack, and R. Kern, "A comparison of layout based bibliographic metadata extraction techniques," in *Proc. 2nd Int. Conf. Web Intell., Mining Semantics (WIMS)*, 2012, p. 19.
[12] R. Habib and M. T. Afzal, "Sections-based bibliographic coupling for research paper recommendation," *Scientometrics*, vol. 119, no. 2, pp. 643–656, May 2019.
[13] S. Eskenazi, P. Gomez-Krämer, and J.-M. Ogier, "A comprehensive survey of mostly textual document segmentation algorithms since 2008," *Pattern Recognit.*, vol. 64, pp. 1–14, Apr. 2017.

[14] S. Mao, A. Rosenfeld, and T. Kanungo, "Document structure analysis algorithms: A literature survey," *Proc. SPIE*, vol. 5010, pp. 197–208, Jan. 2003.
[15] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. D. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko, "Vispubdata.org: A metadata collection about IEEE visualization (VIS) publications," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 9, pp. 2199–2206, Sep. 2016.
[16] S. Marinai and H. Fujisawa, *Machine Learning in Document Analysis and Recognition*, vol. 90. Berlin, Germany: Springer, 2007.
[17] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: A literature survey," *Int. J. Digit. Libraries*, vol. 17, no. 4, pp. 305–338, Nov. 2016.
[18] A. M. Khan, A. Shahid, M. T. Afzal, F. Nazar, F. S. Alotaibi, and K. H. Alyoubi, "SwICS: Section-wise in-text citation score," *IEEE Access*, vol. 7, pp. 137090–137102, 2019.
[19] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*. San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 282–289.
[20] K. Bijari, H. Zare, H. Veisi, and H. Bobarshad, "Memory-enriched big bang–big crunch optimization algorithm for data clustering," *Neural Comput. Appl.*, vol. 29, no. 6, pp. 111–121, Mar. 2018.
[21] I. G. Councill, C. L. Giles, and M.-Y. Kan, "ParsCit: An open-source CRF reference string parsing package," in *Proc. LREC*, vol. 8, 2008, pp. 661–667.
[22] A. Prasad, M. Kaur, and M.-Y. Kan, "Neural ParsCit: A deep learning-based reference string parser," *Int. J. Digit. Libraries*, vol. 19, no. 4, pp. 323–337, Nov. 2018.
[23] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, and Ł. Bolikowski, "CERMINE: Automatic extraction of structured metadata from scientific literature," *Int. J. Document Anal. Recognit.*, vol. 18, no. 4, pp. 317–335, Dec. 2015.
[24] O. R. Berg, S. Oepen, and J. Read, "Towards high-quality text stream extraction from PDF: Technical background to the ACL 2012 contributed task," in *Proc. Special Workshop Rediscovering Years Discoveries*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 98–103.
[25] J. Wu, J. Killian, H. Yang, K. Williams, S. R. Choudhury, S. Tuarob, C. Caragea, and C. L. Giles, "PDFMEF: A multi-entity knowledge extraction framework for scholarly documents and semantic search," in *Proc. 8th Int. Conf. Knowl. Capture*, 2015, p. 13.
[26] A. Constantin, S. Pettifer, and A. Voronkov, "PDFX: Fully-automated PDF-to-XML conversion of scientific literature," in *Proc. ACM Symp. Document Eng.*, 2013, pp. 177–180.
[27] S. Klink and T. Kieninger, "Rule-based document structure understanding with a fuzzy combination of layout and textual features," *Int. J. Document Anal. Recognit.*, vol. 4, no. 1, pp. 18–26, Aug. 2001.
[28] H. Déjean and J.-L. Meunier, "A system for converting PDF documents into structured XML format," in *Proc. Int. Workshop Document Anal. Syst.* Berlin, Germany: Springer, 2006, pp. 129–140.
[29] C. Ramakrishnan, A. Patnia, E. Hovy, and G. A. Burns, "Layout-aware text extraction from full-text PDF of scientific articles," *Source Code Biol. Med.*, vol. 7, no. 1, p. 7, Dec. 2012.
[30] M. Granitzer, M. Hristakeva, K. Jack, and R. Knight, "A comparison of metadata extraction techniques for crowdsourced bibliographic metadata management," in *Proc. 27th Annu. ACM Symp. Appl. Comput. (SAC)*, 2012, pp. 962–964.
[31] D. Tkaczyk, L. Bolikowski, A. Czeczko, and K. Rusek, "A modular metadata extraction system for born-digital articles," in *Proc. 10th IAPR Int. Workshop Document Anal. Syst.*, Mar. 2012, pp. 11–16.
[32] H. H. N. Do, M. K. Chandrasekaran, P. S. Cho, and M. Y. Kan, "Extracting and matching authors and affiliations in scholarly documents," in *Proc. 13th ACM/IEEE-CS Joint Conf. Digit. Libraries (JCDL)*, 2013, pp. 219–228.
[33] T. Kiss and J. Strunk, "Unsupervised multilingual sentence boundary detection," *Comput. Linguistics*, vol. 32, no. 4, pp. 485–525, Dec. 2006.
[34] S. Klampfl, M. Granitzer, K. Jack, and R. Kern, "Unsupervised document structure analysis of digital scientific articles," *Int. J. Digit. Libraries*, vol. 14, nos. 3–4, pp. 83–99, Aug. 2014.
[35] C.-T. Tsai, G. Kundu, and D. Roth, "Concept-based analysis of scientific literature," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 1733–1738.
[36] B. Lowagie, *iText in Action*. Shelter Island, NY, USA: Manning, 2011.
[37] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
[38] A. Dey, "Machine learning algorithms: A review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016.
[39] R. Choudhary and H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," in *Proc. Int. Conf. Mach. Learn. Data Sci. (MLDS)*, Dec. 2017, pp. 37–43.
[40] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2018.
[41] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2009.
[42] M. Abdar, U. R. Acharya, N. Sarrafzadegan, and V. Makarenkov, "NE-nu-SVC: A new nested ensemble clinical decision support system for effective diagnosis of coronary artery disease," *IEEE Access*, vol. 7, pp. 167605–167620, 2019.
[43] M. Claesen, F. De Smet, J. A. K. Suykens, and B. De Moor, "Fast prediction with SVM models containing RBF kernels," 2014, *arXiv:1403.0736*. [Online]. Available: http://arxiv.org/abs/1403.0736
[44] S. Hiregoudar, K. Manjunath, and K. Patil, "A survey: Research summary on neural networks," *Int. J. Res. Eng. Technol.*, vol. 3, no. 15, pp. 385–389, May 2014.
[45] C.-W. Hsu *et al.*, "A practical guide to support vector classification," Nat. Taiwan Univ., Taipei, Taiwan, Tech. Rep., 2003.
[46] A. Dimou, A. Di Iorio, C. Lange, and S. Vahdati, "Semantic publishing challenge—Assessing the quality of scientific output in its ecosystem," in *Semantic Web Evaluation Challenge*. Cham, Switzerland: Springer, 2016, pp. 243–254.

MUHAMMAD WAQAS AHMED received the M.Sc. degree in computer engineering from the Center of Advanced Studies in Engineering, University of Engineering and Technology, Taxila, Pakistan. He is currently pursuing the Ph.D. degree (by research) in the field of information mining, machine learning, and natural language processing with the Center of Distributed and Semantic Computing, Capital University of Science and Technology, Islamabad. He has 12 years of industrial experience in software development, data warehousing, and business intelligence in health care, manufacturing, and oil and gas sectors. He has two years of academic experience in teaching and project supervision.

MUHAMMAD TANVIR AFZAL received the M.Sc. degree in computer science from Quaid-i-Azam University, Islamabad, Pakistan, and the Ph.D. degree (Hons.) in computer science from the Graz University of Technology, Austria. He has been associated with academia and industry at various levels for the last 20 years. He is currently serving as a Professor with the Department of Computer Science, Capital University of Science and Technology, Islamabad. He authored more than 100 research articles in leading peer-reviewed journals and conferences in the field of data science, information retrieval and visualization, semantics, digital libraries, and scientometrics. He has authored two books and has edited two books in computer science. His cumulative impact factor is more than 60, with citations over 500. He played pivotal role in making collaborations between MAJU-JUCS, MAJU-IICM, and TUG-UNIMAS. He conducted more than 100 curricular, co-curricular, and extra-curricular activities in the last five years, including seminars, workshops, national competitions (ExcITeCup), and invited international and national speakers from Google, Oracle, IICM, IFIS, and SEGA Europe. Under his supervision, more than 60 postgraduate students (M.S. and Ph.D.) have defended their research theses successfully and a number of M.S. and Ph.D. students are pursuing their research under his guidance. He received the Gold Medal for his M.Sc. degree. He served as the Ph.D. symposium chair, the session chair, the finance chair, a committee member, and an editor for several IEEE, ACM, Springer, and Elsevier international conferences and journals. He is serving as the Editor-in-Chief for reputed impact factor journal *Journal of Universal Computer Science*.