A Deep Multimodal Investigation To Determine the *Appropriateness* of Scholarly Submissions

Tirthankar Ghosal
Indian Institute of Technology Patna
Patna, Bihar, India
tirthankar.pcs16@iitp.ac.in

Ashish Raj
Indian Institute of Technology Patna
Patna, Bihar, India
ashish.cs15@iitp.ac.in

Asif Ekbal
Indian Institute of Technology Patna
Patna, Bihar, India
asif@iitp.ac.in

Sriparna Saha
Indian Institute of Technology Patna
Patna, Bihar, India
sriparna@iitp.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology Patna
Patna, Bihar, India
pb@iitp.ac.in

ABSTRACT

Present day peer review is a time-consuming process and is still the only gatekeeper of scientific knowledge and wisdom. However, the rapid increase in research article submissions these days across different fields is posing significant challenges to the current system. Hence the incorporation of Artificial Intelligence (AI) techniques to better streamline the existing peer review system is an immediate need in this age of rapid scientific progress. Among many, one particular challenge these days is that the journal editors and conference program chairs are overwhelmed with the ever-increasing rise in article submissions. Studies show that a lot many submissions are not well-informed and do not fit within the scope of the intended journal or conference. Here in this work, we embark on to investigate how an AI could assist the editors and program chairs to identify potential *out-of-scope* submissions based on the past accepted papers of the particular journal or conference. We design a multimodal deep neural architecture and investigate the role of every possible channel of information in a research article (full-text, bibliography, images) to determine its appropriateness to the concerned venue. Our approach does not involve any handcrafted features, solely depends on the past accepting activity of the venue, and thereby achieves significant performance on two real-life datasets. Our findings suggest that a system of this kind is possible and with reasonable accuracy could assist the editors/chairs in flagging out inappropriate submissions.

CCS CONCEPTS

• **Information systems** → **Information systems applications**.

KEYWORDS

peer review, deep learning, multimodality, scope of a journal, appropriateness of a research article

1 INTRODUCTION

Peer Review is the benchmark of modern-day research validation. In spite of having certain inherent flaws like sometimes being biased, time-consuming, arbitrary [13], peer review is still the widely accepted method to document scientific progress. However, with the exponential rise in article submissions, thanks to the *Publish or Perish* syndrome in academia [8], the peer review system is threatened with a never-seen-before information overload [28]. The electronic preprints repository arXiv receives 500-600 new submissions daily with an additional 300-400 submissions update[1]. Editors and Conference Chairs are overwhelmed with the huge number of submissions made[2], and they face a dearth of good reviewers to review the submissions [18]. Sometimes the editors and chairs are left with no other option than to assign papers to novice/out-of-domain researchers or graduate students which often results in poor quality reviews, thus affecting the subsequent decision and the entire academia in general. However, studies [15, 30] show that a good number of submissions are not at all informed ones and sometimes are submitted to wrong venues. Unfortunately, in spite of having merit, some articles do not fit to the aims and scope of the intended venue and have to suffer *Desk-Rejection.*

After submission, the first stage in the academic peer review process commences at the editors' desk, wherein the journal editor decides whether the submitted article fits the aims and scope of the concerned journal. The Editor-in-Chief always looks at the scope of the research study with respect to that of the journal before deciding whether to send it for review. Surprisingly a lot many submissions are rejected at the desk[3] [29] popularly known as Desk-Rejection. It means the editor of the particular journal deems the submitted article unsuitable enough to forward to the expert reviewers for meticulous evaluation. Many reasons account for this activity, foremost being that the submitted article is *out-of-scope* of the intended journal [14, 15]. It may signify that the research findings are of interest to a very narrow or specialised audience that the journal does not cater to specifically. A study on the recently released PeerRead dataset [17] reveals that *Appropriateness* of a manuscript to a certain conference (ACL 2017) is the most correlated aspect with the final recommendation by the reviewers[4].

Our objective here in this work is to reduce this category of information overload and help the editors to identify potential misfit submissions. With the current state of AI, we do not support a fully automated system. Rather we vouch for an editorial assistant

[1] https://blogs.cornell.edu/arxiv/2018/01/19/a-day-in-the-life-of-the-arxiv-admin-team/

[2] Apparently CVPR, NIPS, AAAI 2019 received over 5100, 4900, 7000 submissions respectively!

[3] we show in our earlier work [15] that: to an extent of 50% of desk-rejections account for articles not being *within-scope* of the journal concerned

[4] a positive correlation of as high as 0.49

who could isolate potential *out-of-scope* submissions to be further looked upon by editors/chairs and thereby speed up the review process. We try to imitate the human nature of comprehending a research piece and hence consider all available information within the manuscript. We do not attempt to define the scope of a journal or manually craft features for the same. Instead, we take a pragmatic approach and let our deep neural architecture learn the *domain of operation* of the journal from its accepted papers. In that way, we believe that the deep neural network automatically learns the extensive and different aspects/views of scope for various venues.

We further explore if we could identify the specific venue of a prospective article among venues having overlapping nature of scope. We perform our experiments on a set of real articles curated from six different Computer Science journals (Dataset-I). We curate another dataset (Dataset-II) from open access articles of Artificial Intelligence (AI), Machine Learning (ML), Computer Vision (CV), and Natural Language Processing (NLP) to facilitate our study.

1.1 Why Multimodal Processing?

Research articles are essentially multimodal, especially considering those from STEM disciplines. The variety of figures, graphs complements the text in the article and enables the reader to understand the proposition and analysis better. While images may not always be that significant to certain disciplines, but do play a major role in the comprehension of the research in others (for e.g., natural sciences and medicine [23]). Here in this work, we are intrigued to see if images in research articles contribute to this problem of domain-based research article classification.

1.2 Motivation and Contribution

The motivation behind this work is to efficiently manage the exponential rise in article submissions to journals and conferences these days [20]. The rapid growth in scientific production may threaten the capacity for the scientific community to handle the ever-increasing demand for peer review of scientific publications [18]. In spite of having merit, many papers ($\sim$ 30%) [15, 30] are rejected from the desk simply because they are a misfit to the journals aims, scope, and audience. However unfortunate it is, this phenomena still consume the precious time of all the stakeholders (authors/editors/program chairs and even sometimes reviewers) associated in the peer review pipeline. Thus a system of this kind could eventually assist the journal editors and conference chairs to make better-informed decisions regarding the appropriateness of an article to a submitted venue and quickly locate inappropriate *out-of-scope* submissions. Even potential early-career authors may reap the benefit, and they could be confident about the aptness of their research to the desired journal/conference. This would prove as a huge time-saver for both authors and editors and eventually speed up the overall peer review process. The contributions of the current work are:

- Proposing a multimodal deep neural architecture to classify article submissions based on their aptness to the concerned venue.
- Investigating the role of all possible channels of information in a research article towards the problem. A large scale study was done on six journals and fourteen top-tier conferences of a specific discipline.
- A small step towards an AI-assisted peer review system to cope with the information overload in academia

Good performance over cross domain paper data (Dataset-I) motivated us to go further and investigate the viability of our approach over intradomain data (Dataset-II). We investigate if our proposed method can identify papers belonging to specific sub-domains (*NLP, CV*) of a particular field (*Artificial Intelligence*). Kindly refer to Section 4 for the dataset description. We achieve significant performance improvement over standard baselines. We also show that using paper metadata we could achieve comparable performance as full-text.

We organise the remainder of the paper as follows: in the subsequent section, we review relevant work. In Section 3, we discuss the problem. We introduce the datasets in Section 4 and the experimental setup in Section 6. We propose our architecture in Section 5 and present the results in Section 7. Finally, we conclude with our future directions in Section 8.

2 RELATED WORKS

There had been quite a lot of discussion and work on publication mining, AI in peer review lately. Authors did a thorough study on the various means of computational support to the peer review system in [26]. Reference [22] explored an evolutionary algorithm to improve editorial strategies in peer review. However, to the best of our knowledge, we are the first to explore this problem of article classification under the light of scope identification using deep learning. Automated article classification to predict Accept/Reject decisions is explored in [17]. Our earlier efforts towards the current problem with handcrafted features are documented in [14, 15]. We hand-craft features from several sections of a manuscript that contributes towards determining its scope.

The current work comes close to journal recommendation for academic manuscripts. However, most of the journal recommender systems only consider the *Title* and *Abstract* of the paper for generating a suggestion of potential journals where the author may consider to submit her work. Our problem is a bit different and mostly targeted towards assisting the editors/chairs to let them identify potential *out-of-scope* submissions. We consider every possible channel of information in a research article (text, image, bibliography) to arrive at a decision. Scope Detection as a problem has not yet been studied exclusively in literature. Most of the reputed journal publishers have their systems that suggest relevant journals to an author against her work. Examples could be given of Journal Finder by Elsevier[5], Springer Journal Suggester[6], EDANZ Journal Selector[7],etc. Also some web-services like JANE (Journal/Author Name Estimator)[8] [27], eTBLAST [12], GoPubMed [10], HubMed [11], Pubfinder [16], etc. suggest relevant biomedical literatures from PubMed[9] or MEDLINE[10] databases upon user query (typically the title and abstract of the article for which the user wants to find a

[5] http://journalfinder.elsevier.com/
[6] http://journalsuggester.springer.com/
[7] https://www.edanzediting.com/journal-selector
[8] http://jane.biosemantics.org/
[9] https://www.ncbi.nlm.nih.gov/pubmed/
[10] https://www.nlm.nih.gov/bsd/pmresources.html

suitable journal). These systems mostly rely on domain-specific vocabulary match between the prospective article and different journals to generate a suitable match. Users generally have to submit their article title, abstract and/or keywords to get a list of potential journals where they could submit their article. There had been quite a lot of work on venue recommendation systems for academic manuscripts. Mention may be made of some notable works [1, 3, 5, 21, 32].

Multimodal deep learning from texts, images and videos is a popular NLP problem and is widely explored in the works of [24, 25]. However, to the best of our knowledge, there is hardly any work on multimodal learning in the scholarly text processing domain.

3 SCOPE DETECTION

Submitting a manuscript to an unsuitable journal is one of the most common mistakes committed by authors. Usually, novice/early-career researchers and sometimes even seasoned researchers commit this error. The *scope of a journal* is a very broad term and vary across different journals[11]. We enlist some of our observations from the study of Desk-Rejected due to Out-Of-Scope (DR-OOS) articles. Special thanks to our academic collaborator Elsevier, to support this investigation with necessary resources.

3.1 Desk-Rejection Observations

- If one submits a paper from Computer Networks to an Artificial Intelligence journal; it is out-of-scope. However, naive as it may sound, this activity not rare, ultimately resulting in desk-rejection.
- Again a paper which is too specific to a particular domain of interest (e.g., Neural Networks) sometimes is not accepted by a journal which caters to a broader perspective (e.g., Artificial Intelligence).
- Similarly, a journal which accepts review papers (e.g., ACM Computing Surveys) may not consider a method paper and vice-versa.
- A theoretical journal (e.g., Theoretical Computer Science) would not be interested in an application-focused paper even though the domain may be identical.
- Sometimes the scope is also linked to the quality of the manuscript. A journal may cater to a vast area of topics but only looks for high quality, original and innovative submissions (for example Nature or Science) which have the potential to induce a significant impact post-publication.
- Again we observe that *scope* of a journal is *time-variant* and usually gets streamlined over time. This behaviour reflects the advancements in science and popularity of topics in the scientific community (for e.g., Deep Learning is hugely popular now in NLP, AI, and CV community).

Most of the journals ask the potential authors to go through the past accepted papers of that journal to get a feel of the type of papers they publish and the audience they cater to. The past publishing activity of a journal defines its *domain of operation* and the topics it is interested in. However, the problem of misinformed submissions is still glaring at the present-day peer review system; authors do make less-informed choices, resulting in wastage of precious time of both the authors and journal editors.

3.2 Scope of a Journal

Scope, simply stated, is the journal's purpose or objective. It is what the publication wants to achieve by delivering its content to the readers. The relevance/similarity of an article with published papers is a good indicator of its domain. However, the article should not be that similar so that it falls short of the originality/novelty criteria. The domain of a journal is one variant of its scope. In spite of having merit, many submissions face rejections because they do not fit to the declared domains of the journal. So we understand that *Scope* of a journal is very subjective and is hard to define in quantitative terms. There are many views, and we could aptly cast it as a multiview problem. However, in this work, we attempt to explore a limited definition of journal scope: *the domain or range of topics a given journal caters to.* Our experience with the study of desk-rejected papers reveals that out-of-domain submissions are common and account for a large number of desk rejections. Here we try to understand which section of the manuscript contributes more to define its domain and belongingness to a particular journal. However, this in no way mitigates the broader perspective of scope we discussed earlier. Even the available journal recommender systems check the *domainness* of a manuscript to its published articles by simple content words match. *Accepted published articles are thus the benchmark of reference.*

3.3 Task 1

We model the problem as a binary classification one: *classifying a given article into* ***within-scope*** *or* ***out-of-scope*** *classes.* We train separate models on accepted and *out-of-scope* articles of each venue to test the suitability of an incoming article to the scope of the particular journal/conference.

3.4 Nature of Scope

The scope of a venue is not constant. It changes with time with the progress in scientific knowledge. Even many venues have an overlapping domain of operation. For e.g., Artificial Intelligence (AI), Machine Learning (ML) techniques finds applications in Natural Language Processing (NLP) or Computer Vision (CV) problems. So a certain paper with ML techniques applied to an NLP problem may seem to qualify for both NLP and ML venues. NLP and ML are sub-fields of AI. The distinction in the topics of interest for such cases is not very pronounced. They would share a similar kind of vocabulary, citations, techniques, named-entities, and, even authors. However, there are some subtle differences in the motivations, aims of such venues which define their scope, their *domain of operation.* Whereas the focus of the ML venue would be towards finding some novelty in the ML techniques used, the NLP venue would look for novelty in the problem and in the corresponding approach towards the solution. However, it is still not always distinctive given the interdisciplinary nature of research.

3.5 Task 2

Here we are interested to see if we can predict the actual venue of an article among potential venues with a nearly identical domain

[11] https://wordvice.com/choosing-the-right-journal-scope-issues/

of operation. With this motivation, we design our second set of experiments on Dataset-II described in Section 4.2. With Dataset-II we want to go deeper and see how the various channels of information contribute to identifying the class of a research article which may belong to multiple venues having overlapping nature of scope. We model the problem to handle multi-class scenarios where the objective would be: *To which venue a particular paper should go when there are multiple potential venues?* We seek how past accepted papers in these venues having overlapping nature of scope could effectively identify the place holder of a new submission? This could be effectively seen from the viewpoint of a prospective author and towards a venue recommender system.

3.6 Formalizing the problem

Given a set of N research articles, the objective is to minimize the negative log likelihood over the classes:

$$-\sum_{c=1}^{M} y_{o,c} log(p_{o,c})$$

where $p_{o,c} = f(o(x_n))$, o is the network's output, x_n is the multichannel multi-modal input and $y_{o,c}$ is the indicator if class label c is the correct classification for observation o. In our case, $M = 2$ and $f = sigmoid$ for Task 1 (binary classification) and $M = 3$ and $f = softmax$ for Task 2 (multi-class classification). Here the modality is two: text and image. Further the text modality has two distinct channels: full text and bibliography.

4 DATA DESCRIPTION AND ANALYSIS

Getting hold of actual desk-rejected data is hard due to proprietary and confidentiality reasons. However, we curate two datasets to proceed with our experiments. One we got from our collaborator (Dataset-I) and the other we create from open access articles (Dataset-II). The motivation behind experimenting with these two datasets are slightly different. While with Dataset-I we cater to the need at the editors' end, with Dataset-II we find an author perspective to the problem. We follow an 80

4.1 Dataset-I

We create Dataset-I with the papers from the following six Computer Science journals: Artificial Intelligence (ARTINT), Computer Networks (COMNET), Journal of Computer Network and Applications (JNCA), Computer Standards and Interfaces (CSI), Simulation Modelling Practice and Theory (SIMPAT), Statistics and Probability Letters (STATPRO). We are thankful to our collaborator Elsevier, for providing us with a subset of desk-rejected data of these six journals. In our earlier study [14], we show that nearly 50% of desk-rejections accounts for articles not being within scope. However, for a deep learning experimental setup, the actual available *out-of-scope* papers were not sufficient. Hence, along with with actual *out-of-scope* instances from the desk-rejected articles, we also select articles from other journals to serve as the negative instances for a given journal. The intuition is simple: *Accepted articles of other remotely related journals would be out-of-scope of the current journal under study.* We consider accepted papers from a set of 17 different Computer Science journals to simulate our negative data. This we do to make our negative data as diverse as possible. We had all the accepted articles of the six journals as our positive data. Table 1 illustrates our Dataset-I statistics. The total number of images and bibliography items for each journal speaks high of the volume of information they carry within the manuscripts.

ARTINT journal invites original research in theory, techniques and applications of Artificial Intelligence. The domain is vast. COMNET is for topics on Computer Networks and is somewhat restricted in the area as compared to ARTINT. JNCA is close to the scope of COMNET and has overlapping *topics of interest.* The journal Simulation Modelling Practice and Theory (SIMPAT) provides a forum for original, high-quality papers dealing with any aspect of systems simulation and modelling. Computer Standards and Interfaces (CSI) focusses on quality of software, well-defined interfaces (hardware and software), the process of digitalisation, and accepted standards in these fields. STATPRO is all about Statistical and Probability theories and has a limited scope as compared to others.

4.2 Dataset-II

Dataset-II comprises of open access articles from several top-tier conferences in the field of Artificial Intelligence and Machine Learning, Natural Language Processing, and Computer Vision. NLP and CV are sub-fields of AI and are currently heavily reliant on ML techniques. Hence there is an overlapping domain of interest between AI and NLP/CV conferences. AI conferences accept papers that address challenges in both NLP and CV. However, AI conferences also cater to several other areas like Robotics, Data Mining, Knowledge Discovery, Machine Learning, etc. With the recent interest and rapid progress in AI/ML domain, every other STEM discipline is using AI/ML, thus making the scope of AI very broad. However, there are some subtle distinctions in aims and motivations behind general AI conferences and more specific venues from NLP and CV. Certain domain-specific papers in NLP and CV would be of more interest to a specialist audience than a general one. Hence with this dataset, we explore, to which conference category a particular paper should belong? With our earlier dataset, the distinctions were pretty obvious while with Dataset-II certainly there is an overlap in the *domain of operation* of the venues. We investigate how our deep network trained on previously accepted papers of those allied venues could correctly identify the suitable venue of a new submission. Although we perform a 3-class classification, our model could be suitably tuned to handle multi-class scenarios. The distal objective is to build a recommender system which could efficiently guide the authors to consider a more suitable venue for their manuscripts. We also explore the effects of different modalities in different categories (NLP, CV) of the same domain (AI).

Table 2 shows the data statistics for Dataset-II. For AI/ML, we consider papers from International Conference on Learning Representations (ICLR), Association for the Advancement of Artificial Intelligence (AAAI) Conference on AI, International Joint Conference on Artificial Intelligence (IJCAI), and NeurIPS (Conference on Neural Information and Processing Systems, previously called NIPS). For NLP, we take papers from Association for Computational Linguistics (ACL), North American Association for Computational Linguistics (NAACL), European Association for Computational Linguistics (EACL), Conference on Empirical Methods in Natural

Journals	Accepted		Rejected		Actual Negative		#Images		#Bib. Entries		# FT Sentences	
	# Train	# Test	# Train	# Test	# Train	# Test	# Train	# Test	# Train	# Test	# Train	# Test
ARTINT	1348	337	1239	308	270	68	5660	1518	44379	11174	725791	174611
COMNET	2957	740	2934	729	365	92	6878	2685	64613	16602	1017532	261342
STATPRO	4345	1087	3956	981	307	77	1734	910	25598	6725	646893	160351
JNCA	1614	404	1450	365	24	6	14923	3705	45958	10629	938470	234754
SIMPAT	1228	307	1149	285	419	103	7850	4093	24454	6222	325053	86010
CSI	1663	416	1499	375	17	5	4532	1303	16700	4748	287150	76769

Table 1: Dataset-I Statistics (Elsevier), FT→Full-Text, Actual Negative are the instances (papers) which were desk-rejected due to *out-of-scope* from the concerned journal, Bib→Bibliography

Category	Conferences	#Images		#Bibliography		#Sentences		#Papers	
		# Train	# Test	# Train	# Test	# Train	# Test	# Train	# Test
AI/ML	**IJCAI, AAAI, ICLR, ICML, NIPS**	6596	3169	163642	29011	1324259	200424	6719	932
CV	**CVPR, ICCV, ECCV**	7290	4804	191943	41413	1209511	223876	5403	1011
NLP/CL	**ACL, NAACL, EACL, COLING, CoNLL, EMNLP**	15200	2666	165345	29456	1193096	190613	5842	920

Table 2: Dataset-II Statistics (Open Access AI/ML/NLP/CV Papers), This statistics signify the volume of information processing corresponding to the three modalities

Language Processing (EMNLP), International Conference on Computational Linguistics (COLING), and Conference on Natural Language Learning (CoNLL). For Computer Vision, we consider papers from The Conference on Computer Vision and Pattern Recognition (CVPR), European Conference on Computer Vision (ECCV), and International Conference on Computer Vision (ICCV). Since the rapid progress of Deep Learning in AI has its origin in the ImageNet competition in 2012 [19], we consider papers from these venues from 2012 till 2018.

4.3 Pre-processing

The original articles are in PDF. We use the Science Parse library [12] to convert the PDF into. JSON encoded files for information extraction. Tables, Formulas are distorted in the process, and we exclude those from further processing. We extract figures from the raw PDF's using the PDFFigures 2.0 library [6]. We extract the bibliography section and consider only the citation titles and venues in our experiments. Paper titles and venues contain certain domain-specific vocabulary and are a good indicator of the domain of the paper [15]. The other elements in the bibliography (Authors, Year, Page Numbers, Publisher, etc.) has little relevance to our task, and so we ignore them. We remove stop words and certain common words (for e.g., *International, Journal, Conference, Proceedings, etc.*) from the citations. We create a vocabulary list from citation titles and venues and use it in the Bag-of-Words (BoW) model discussed in Section 5.3.

5 METHODOLOGY

We choose to investigate a deep neural solution to this problem because the definition of scope is not invariant across journals/conferences (discussed in Section 3.1). Our idea is to let the network learn the *scope* of a venue from its past accepted articles. We present the overall architecture in Figure 1. We present the hyperparameter details in Section 6.2. Our architecture is divided into two phases. In Phase-I we learn the feature representation from various modalities. In Phase-II we learn the importance of the modalities via attention mechanism, weigh them accordingly, fuse them, and finally classify the article into *Within Scope* or *Out-of-Scope.*

Phase I: Representation Learning of Multimodal Paper Features

Here we learn useful features from different paper components (Full-Text, Images, Bibliography).

5.1 Textual Feature Extraction

We extract full-text sentences from each research article and use the Transformer variant of the Universal Sentence Encoder (USE) [4] to encode the full-text sentences into 512 dimensional semantic vectors. We then stack the sentence vectors to form the document representation. Next we train our *Textual Modality Feature Extractor* by passing this document representation through an end-to-end Bi-Directional Long Short Term Memory (Bi-LSTM) network followed by a Multi-Layer Perceptron (MLP-1) with a final sigmoid layer for classification. We use the activations of the preceding fully-connected layer of MLP-1 as the document-level feature representation of Text T.

[12] https://github.com/allenai/science-parse

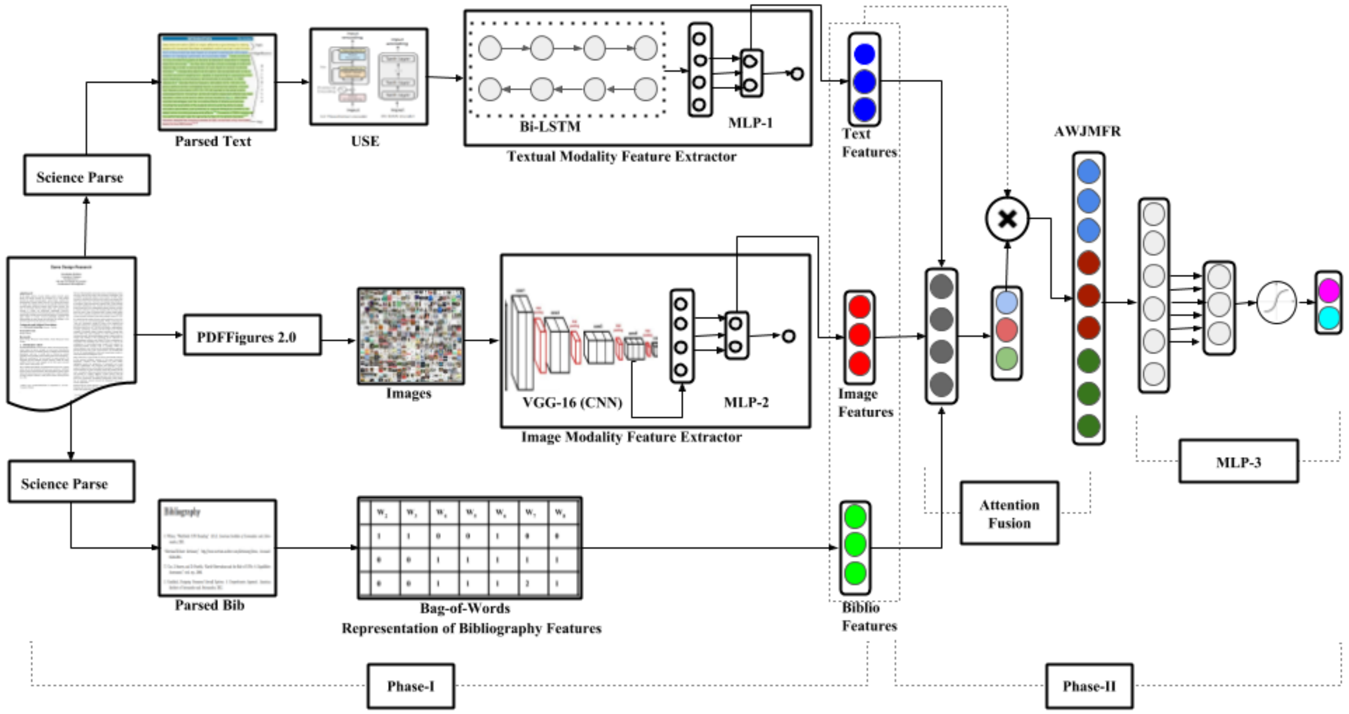

Figure 1: Proposed Deep Multimodal Neural Architecture for Scope Detection

Let $[S_i]$ be the output sentence representation of the Universal Sentence Encoder. We use separate LSTM modules to produce forward and backward hidden vectors, which are then concatenated:

$$\overrightarrow{h_t} = \overrightarrow{LSTM}_t([S_i])$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}_t([S_i])$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$$

We pass the final hidden layer of the Bi-LSTM to a multi-layer perceptron (MLP) to obtain the final representation vector (V_T) of the paper text.

$$V_T = f_{mlp}([h_t]; \theta_{mlp})$$

where f_{mlp} denotes a three-layer-MLP, and θ_{mlp} denotes the parameters in it.

5.2 Image Feature Extraction

First we extract the figures from each paper using PDFFigures 2.0 [7]. Then we make use of the pre-trained VGG-16 with ImageNet [9] weights to train our *Image Modality Feature Extractor*. The *Image Modality Feature Extractor* consists of an end-to-end 16-layer deep Convolutional Neural Network (CNN) followed by a Multi-Layer Perceptron (MLP-2) network. We freeze the first seven layers of the deep VGG-16 CNN and make the subsequent nine layers trainable. The output of the final affine layer of the VGG-16 CNN is the input to MLP-2 which has 3-layers with a final sigmoid layer for classification. Like previous, we use the activations of the preceding fully-connected layer of MLP-2 as the features of Image I. Hence,

$$V_I = f_{mlp}([h_{CNN}]; \theta_{mlp})$$

Where V_I is the final image representation, h_{CNN} are the activations of the last hidden layer of VGG-16 CNN, f_{mlp} denotes a three-layer-MLP, and θ_{mlp} denotes the parameters in it. We concatenate all the image representations for a paper to generate the final image representation. If no images are there in a paper, we use a zero-padded vector of dimension equivalent to 8 images as a feature vector for image modality.

We also take the *Bag-of-Words* representation of the **image captions** and fuse it with the corresponding image feature representation via concatenation.

Why training with VGG-16?
The number of images found in the research papers is not always adequate to train a deep neural feature extractor from scratch. VGG-16 is a deep CNN with 16 layers trained on millions of images. VGG-16 is also a *state-of-the-art* object detector and has been used for transfer learning in many use-cases. Hence we use pre-trained VGG-16 to aid in our high-level feature extraction from the paper images. We freeze the first seven layers as they usually discover low-level features like edges etc.

5.3 Bibliography Feature Extraction

In an earlier work [15] we show that the Bibliography section consists of important domain information regarding the scope of an article to a venue. Especially the citation titles and venues hold significant domain information. Hence we consider *Bibliography* as a separate channel of the text modality here. We find that the vocabulary size of citations for a particular venue (journal/conference) is limited. Hence we proceed with a simple *Bag-of-Words* model to generate the bibliographic feature representations for this channel. To obtain bibliographic feature representation vector V_B of the document, we concatenate the BoW vectors of bibliographic paper titles (t_i) and venues (v_i).

$$V_B = BoW(v_i)||BoW(t_i)$$

Phase-II: Attention Weighted Multimodal Classification

5.4 Attention-Based Multimodal Fusion

At this stage we have the feature representations from the three modalities (Full-Text, Image, and Bibliography) [13]. To get the best out of each modality, we make use of *Attention* mechanism [2] popular in deep neural networks. Attention mechanism has the ability to focus on the most important parts of an object relevant to the classification, improving the performance of the baseline deep neural networks. The attention mechanism has been successfully employed in several NLP tasks such as sentiment analysis [31]. The motivation behind using the attention layer is that: *Not all modalities contribute equally to determine the domain of a research article pertaining to a certain venue.* To prioritise only important modalities, we use an attention layer, which takes as an input feature representations from the text, image, and bibliography modalities and outputs an attention score for each modality. Using these scores, the modality contributing more would have higher attention weights. We take the dot product of the respective attention weights with the modality representations and fuse them via concatenation to form the *Attention Weighted Joint Multimodal Feature Representation (AWJMFR).* The fused multimodal vector F is computed as follows:

Let M_I, M_T, M_B be the feature representation from various modalities where M_I=V_I, M_T=V_T, and M_B=V_B respectively. The dimensions of M_I, M_T, M_B are d_I, d_T, d_B respectively.

$$M = [M_I, M_T, M_B]$$

$$X = ReLU(W_1^T M)$$

$$A = Softmax(W_2^T X)$$

Where W_1 and W_2 are the weights of the first and second layer neurons respectively.
Let the attention weights obtained from the three modalities be

$$A = [A_I, A_T, A_B]$$

We concatenate the modality vectors after scaling them with attention weights and obtain the final feature fusion AWJMFR:

$$F = AWJMFR = [M_I A_I || M_T A_T || M_B A_B]$$

where || signifies concatenation.

5.5 Scope Classification

Finally we pass the AWJMFR through a 3-layer MLP for classification into two classes. We keep *Sigmoid* activation in the final layer as the first task is a binary classification one. For the multi-class problem (Task 2) we keep *Softmax* in the final layer for classification into 3-classes.

6 EXPERIMENTAL SETUP

We discuss the experimental setup in this section.

6.1 Baselines

To the best of our knowledge, there are no works till date which addresses this problem of multimodal research article classification in the scholarly domain. Hence, we keep the unimodal features (Only Text, Only Image, Only Bibliography) as the baselines for our experiments. Majority of the available journal recommender systems takes *Title* and *Abstract* of a paper as input to suggest a relevant journal. Although our objective is not a recommendation, we also investigate the contributions of individual sections to identify the scope of a candidate paper to a journal.

6.2 Hyperparameter Details

We enlist the hyperparameter details in accordance with Figure 1. The end-to-end trainable Bi-Directional LSTM+MLP in the *Textual Modality Feature Extractor* takes input from the Universal Sentence Encoder. Each sentence has dimension 512 and for each paper we set the number of sentences as 500. The batch size is 64 with binary cross entropy as loss function and *Adam* as the optimizer. The activations in the dense layer is *ReLU* whereas the activation in the final MLP-1 layer is *Sigmoid* for binary classification. We ran 10 epochs untill convergence with a learning rate 0.001 and a dropout of 0.3 in MLP-1. Both the Bi-LSTM and MLP-1 has 3 layers. The output of the full-text feature extractor is a representation of 4000 dimension.

The *Image Modality Feature Extractor* comprises of the VGG-16 CNN followed by a 3-layer Multi-layer Perceptron (MLP-2). The input image has dimension 256× 256. We freeze the first 7 layers of pre-trained VGG-16, train the remaining nine layers with the input paper images (set to a maximum of 8 per paper). We take the activations of the affine layer as input to MLP-2. The MLP has *ReLU* activations in dense layer and *Sigmoid* activation in the final layer for classification. The batch size is 128 with binary cross entropy loss and *Adam* optimizer. We continue till ten epochs until convergence with a 0.5 dropout. The output of the *Image Modality Feature Extractor* is a joint representation of images and corresponding captions with dimension 4096×8+|d| where *d* is the length of the image caption vocabulary.

For our Bibliography modality, we follow the simple *Bag-of-Words* model for representing citation title and citation venue. We prune stop words, words with a frequency less than 3 for titles, and less than 6 for venues. The dimension of the output bibliography feature representation depends on the length of the vocabulary for each venue.

For the *Attention* layers in Phase-II of our architecture, we use *ReLU* activations in the dense layer with a dropout of 0.25 and *softmax* in the final layer to learn the attention weights. Further, we

[13] Although a channel of text modality, we consider Bibliography as a separate modality as the text-form in the bibliography is quite different from that in paper body

Journals	**JNCA**		**ARTINT**		**COMNET**		**SIMPAT**		**STATPRO**		**CSI**	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc
Only Title	0.82	84%	0.78	79%	0.77	78%	0.73	73%	0.79	79%	0.77	78%
Only Abstract	0.82	81%	0.87	86%	0.89	88%	0.79	79%	0.88	88%	0.84	86%
Only Image	0.73	74%	0.53	55%	0.37	50%	0.63	64%	0.34	53%	0.57	57%
Image Captions	0.77	76%	0.63	65%	0.82	81%	0.71	70%	0.69	72%	0.67	68%
Full Text	0.93	89%	0.93	93%	**0.96**	**95%**	0.88	88%	**0.93**	**93%**	0.91	93%
Bibliography	0.87	86%	0.83	86%	0.85	84%	0.71	72%	0.84	85%	0.83	83%
Image+Abstract	0.85	86%	0.89	88%	0.88	88%	0.81	80%	0.82	83%	0.85	86%
Image+Full-Text	0.93	92%	0.93	**94%**	0.95	**95%**	0.88	**90%**	0.85	86%	0.92	91%
Image+Bibliography	0.92	90%	0.89	89%	0.86	86%	0.79	81%	0.85	85%	0.85	86%
Image+Full-Text+Bibliography	**0.94**	**95%**	**0.95**	**94%**	0.93	**95%**	**0.89**	**90%**	0.92	**93%**	**0.93**	**94%**

Table 3: Scope Detection (Binary Classification) Results on Dataset-I (Elsevier Journals)

use binary cross-entropy as the loss function and *Adam* optimizer with batch size=64 and 20 epochs. The final layer has *Sigmoid* activations for binary classification into *Within Scope* and *Out of Scope*. For *Task 2*, the final layer has *Softmax* activation with categorical cross-entropy as the loss function.

7 RESULTS AND ANALYSIS

We discuss and analyse our results on the two datasets in the subsequent section.

7.1 Results on Dataset-I

Table 3 shows our experimental results on Dataset-I. Here we want to see if our deep neural network can identify *within Scope* and *Out-of-Scope* papers and test it on a dataset comprising six different Computer Science journals. Different modality and section combinations allow us to understand the significance of each modality/section. This also serves as a means of our ablation study.

7.1.1 ***Image Modality.*** The image modality performs the worst across all the journals. We study the data and find that most of the extracted images are curves/graphs which are generic to all the journals. A major section of those graphical figures is white spaces signifying no object as such. Hence our feature extractor could not discover useful distinguishing features. However, the image+bibliography channel attains a gain of 4%, 3%, 2%, 9%, and 3% over only bibliography in terms of accuracy for JNCA, ARTINT, COMNET, SIMPAT, and CSI respectively. Quite obvious that Statistics and Probability Letters (STATPRO) do not contain enough images and hence image features are not useful here (we observe a significant drop in F-Score values when combined with other channels).

However, we argue that the role of images as a differentiator could be more significant for certain biological, natural science, medicine journals where images featuring real-life objects are more pronounced, present in case studies and form a central part of the research.

7.1.2 ***Bibliography Channel.*** We observe that the bibliography channel achieves comparable performance with the *Only Abstract* input. Where sometimes paper abstracts are not sufficient, the bibliography may come to the rescue. Bibliography section holds a good amount of domain information in citation titles and venues which we exploit in our experiments.

Journals	**AI/ML**		**CV**		**NLP**	
	F_1	Acc	F_1	Acc	F_1	Acc
Only Title	0.75	74%	0.79	80%	0.85	84%
Only Abstract	0.76	71%	0.83	84%	0.87	90%
Only Image	0.75	70%	0.62	67%	0.79	75%
Image Captions	0.65	52%	0.75	78%	0.68	65%
Full Text	0.92	93%	0.92	91%	0.93	93%
Bibliography	0.87	85%	0.90	91%	0.92	94%
Img+Abs	0.95	**95%**	0.91	92%	0.92	92%
Img+FT	**0.96**	**95%**	0.93	92%	0.93	**96%**
Img+Bib	0.86	83%	0.88	92%	**0.94**	95%
Img+FT+Bib	**0.96**	**95%**	**0.94**	**93%**	**0.94**	93%

Table 4: Results on Dataset-II (AI/ML/NLP/CV). Multi-class classification.

7.1.3 ***Text Modality.*** With the sheer volume of information, *Full-Text* is the clear winner, sometimes even better than other modalities combined. When coupled with additional Bibliography and Image information, we observe a gain of 6% (JNCA), 1% (ARTINT), 2% (SIMPAT), 1% (CSI) in terms of accuracy. For COMNET and STATPRO, there is no change. Our attention module emphasised the full-text modality with much higher weights than others. Full-text processing might be computationally expensive, but always there is this trade-off between high accuracy and volume of information processing.

Our best performing model combines all the modalities of information and achieves significant performance improvement over the baselines and individual channels. We observe a gain of more than 10% over individual channels across all journals. Automatically identifying seemingly *out-of-scope* articles is a very crucial yet delicate task. Hence designing a highly accurate system is the need of the problem. Our results clearly suggest that it is required to consider all modalities of information to achieve that goal.

7.2 Results on Dataset-II

Table 4 shows our results on the Dataset-II. Here we can observe identical behavior as in Dataset-I, almost resonating with the earlier findings. Although the objective of the task is a bit different than with Dataset-I (as discussed in Section 3), we still achieve good performance with our overall model. Basically we try to address, among probable venues, to which venue should a prospective paper go?

7.2.1 *Text Modality*.

The most contributing modality is again the Full-Text which is quite obvious. Majority of the recent AI, NLP, and CV papers are Machine Learning/Deep Learning based. So most of the technical aspects are very close. For e.g., Convolutional Neural Network (CNN) was widely used for image processing and Computer Vision problems, but recently has shown great success in dealing with NLP problems. Similarily, we can see several other Machine Learning concepts finding foray in NLP and CV papers. Still the scope of those papers could be differentiated with the problem they address, the data they work on, and the insights they derive.

7.2.2 *Bibliography Channel*.

However, we see that Bibliography channel fares almost close to the Full-Text modality. This is because the type of citations for NLP and CV would be different. AI/ML papers are based on core mathematical and theoretical groundings, many are from disciplines other than NLP, CV; hence have a different category of bibliographic citations in comparsion to more application oriented NLP and CV papers.

7.2.3 *Image Modality*.

Image modality features alone do not perform well. But when augmented with the paper abstract gains an accuracy of more than 14% (at least). Even combination of Bibliography channel with Image features reaches a competitive benchmark as Full-Text. The less performance of images is because images present in papers are not uniform in terms of numbers, quality, etc. Many of them are graphs which convey little information about the domain of the manuscript, as we discuss earlier too.

7.3 Error Analysis

Although few, errors in our system are due to:

1. We were not able to process the text crisply as is there in the paper. Parsing errors, lot of *out-of-vocabulary* words are few reasons. We should have used embeddings generated from scholarly data.
2. Majority of the images (graphs) were similar for all the classes. Less amount of distinctive images.
3. Overlapping nature of textual content (in case of Dataset-II). For e.g., similar technologies used in NLP, CV papers. At least considering the surface form of the texts.

8 CONCLUSION AND FUTURE WORKS

Here in this work, we conduct a thorough study of the role of different modalities and information channels for determining the belongingness of an article to a venue. To the best of our knowledge, we are the first to employ a deep neural network for the problem in hand. Our extensive experiments on an array of journals and conferences show that a highly accurate system of this kind is possible. The definition of scope for each venue is different and is not always dependant on specific topics of interest. Our network learns the scope characteristics of each venue and corresponding *domain of operation* from past accepted papers. With our experiments on the Dataset-II, we are able to address the place holder of a manuscript in highly related venues. This research could be suitably moulded to build a venue recommender system for the authors as well. For the editors, it would be much easier to identify potential misfit submissions and intimate the authors quickly thus accelerating the overall peer review system. The associated codes for this work can be found here [14].

As our future work we would like to concentrate next on:

- Experimenting with journals where images are much more significant (Natural Sciences, Biology, and Medicine).
- Venues which have overlapping nature of scope (Conferences or Journals having the same domain of interest, for e.g., within AI/ML conferences)
- A recommender system for the authors to choose venues wisely. A graded ranking scale of appropriate venues instead of a binary decision.
- Training our network with images from respective journals/conferences only, from topically similar images crawled from the web
- Identifying less relevant images (e.g., graphs) and pruning them out in the workflow
- Addressing the observations in Section 3.1

ACKNOWLEDGMENTS

The first author is supported by Visvesvaraya Ph.D. Scheme of Ministry of Electronics and Information Technology (MeitY), Government of India. The third author, Asif Ekbal gratefully acknowledge the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this work. We thank Elsevier, our academic collaborator, to provide us with requisite data for this study. We also thank the anonymous reviewers for their valuable inputs to improve upon the initial version of the paper.

[14] https://github.com/araj231996/JCDL-19

REFERENCES

[1] Hamed Alhoori and Richard Furuta. 2017. Recommendation of scholarly venues based on dynamic user interests. *Journal of Informetrics* 11, 2 (2017), 553–563.

[2] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. 2016. End-to-end attention-based large vocabulary speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 4945–4949.

[3] Imen Boukhris and Raouia Ayachi. 2014. A novel personalized academic venue hybrid recommender. In *Computational Intelligence and Informatics (CINTI), 2014 IEEE 15th International Symposium on*. IEEE, 465–470.

[4] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*. 169–174. https://aclanthology.info/papers/D18-2029/d18-2029

[5] Zhen Chen, Feng Xia, Huizhen Jiang, Haifeng Liu, and Jun Zhang. 2015. AVER: Random walk based academic venue recommendation. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 579–584.

[6] Christopher Clark and Santosh Divvala. 2016. PDFFigures 2.0: Mining figures from research papers. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*. IEEE, 143–152.

[7] Christopher Andreas Clark and Santosh Kumar Divvala. 2016. PDFFigures 2.0: Mining Figures from Research Papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, JCDL 2016, Newark, NJ, USA, June 19 - 23, 2016*. 143–152. https://doi.org/10.1145/2910896.2910904

[8] Mark De Rond and Alan N Miller. 2005. Publish or perish: bane or boon of academic life? *Journal of Management Inquiry* 14, 4 (2005), 321–329.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255.

[10] Andreas Doms and Michael Schroeder. 2005. GoPubMed: exploring PubMed with the gene ontology. *Nucleic acids research* 33, suppl 2 (2005), W783–W786.

[11] Alfred D Eaton. 2006. HubMed: a web-based biomedical literature search interface. *Nucleic acids research* 34, suppl 2 (2006), W745–W747.

[12] Mounir Errami, Jonathan D Wren, Justin M Hicks, and Harold R Garner. 2007. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic acids research* 35, suppl 2 (2007), W12–W15.

[13] Olivier François. 2015. Arbitrariness of peer review: A Bayesian analysis of the NIPS experiment. *arXiv preprint arXiv:1507.06411* (2015).

[14] Tirthankar Ghosal, Ravi Sonam, Sriparna Saha, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Investigating Domain Features For Scope Detection and Classification of Scientific Articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (7-12). European Language Resources Association (ELRA), Paris, France.

[15] Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, Sriparna Saha, and Pushpak Bhattacharyya. 2018. Investigating Impact Features in Editorial Pre-Screening of Research Papers. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018, Fort Worth, TX, USA, June 03-07, 2018*. 333–334. https://doi.org/10.1145/3197026.3203910

[16] Thomas Goetz and Claus-Wilhelm von der Lieth. 2005. PubFinder: a tool for improving retrieval rate of relevant PubMed abstracts. *Nucleic acids research* 33, suppl 2 (2005), W774–W778.

[17] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard H. Hovy, and Roy Schwartz. 2018. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*. 1647–1661. https://aclanthology.info/papers/N18-1149/n18-1149

[18] Michail Kovanis, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. 2016. The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS One* 11, 11 (2016), e0166387.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[20] Seth S Leopold. 2015. Increased manuscript submissions prompt journals to make hard choices.

[21] Hiep Luong, Tin Huynh, Susan Gauch, Loc Do, and Kiem Hoang. 2012. Publication venue recommendation using author networkâĂŹs publication history. In *Asian Conference on Intelligent Information and Database Systems*. Springer, 426–435.

[22] Maciej J Mrowinski, Piotr Fronczak, Agata Fronczak, Marcel Ausloos, and Olgica Nedic. 2017. Artificial intelligence in peer review: How can evolutionary computation support journal editors? *PloS one* 12, 9 (2017), e0184711.

[23] Henning Müller, Antonio Foncubierta-Rodriguez, Chang Lin, and Ivan Eggel. 2013. Determining the importance of figures in journal articles to find representative images. In *SPIE Proceedings*, Vol. 8674. 9.

[24] Soujanya Poria, Erik Cambria, Amir Hussain, and Guang-Bin Huang. 2015. Towards an intelligent framework for multimodal affective data analysis. *Neural Networks* 63 (2015), 104–116.

[25] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 439–448.

[26] Simon Price and Peter A. Flach. 2017. Computational support for academic peer review: a perspective from artificial intelligence. *Commun. ACM* 60, 3 (2017), 70–79. https://doi.org/10.1145/2979672

[27] Martijn J Schuemie and Jan A Kors. 2008. Jane: suggesting journals, finding experts. *Bioinformatics* 24, 5 (2008), 727–728.

[28] Richard Smith. 2010. Strategies for coping with information overload.

[29] HervÃľ Stolowy. 2017. Letter from the Editor: Why Are Papers Desk Rejected at European Accounting Review? *European Accounting Review* 26, 3 (2017), 411–418. https://doi.org/10.1080/09638180.2017.1347360 arXiv:https://doi.org/10.1080/09638180.2017.1347360

[30] Susan Trumbore, Mary-Elena Carr, and Sara Mikaloff-Fletcher. 2015. Criteria for rejection of papers without review. *Global Biogeochemical Cycles* 29, 8 (2015), 1123–1123.

[31] Yequan Wang, Minlie Huang, Li Zhao, et al. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 606–615.

[32] Shuo Yu, Jiaying Liu, Zhuo Yang, Zhen Chen, Huizhen Jiang, Amr Tolba, and Feng Xia. 2018. PAVE: Personalized Academic Venue recommendation Exploiting co-publication networks. *Journal of Network and Computer Applications* 104 (2018), 38–47.