Exploring the Implications of Artificial Intelligence in Various Aspects of Scholarly Peer Review

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

ABSTRACT

This doctoral work aims to explore the possible avenues in academic peer review where Artificial Intelligence (AI) could have a positive impact. Present day peer review is suffering from many shortcomings, foremost being that it is a very time-delayed process. We identify three potential factors: *Novelty, Scope and Quality* which are central to the theme of scholarly communication process and seek to investigate various techniques encompassing Natural Language Processing, Information Extraction and Retrieval, Data Mining, Bibliographic Analysis, etc. to address those. The objective is to develop an AI assisted decision support system for the editors, reviewers as well as the authors to get preliminary meta information about a prospective manuscript which may aid in decision making and hence speed-up the overall process. The project is challenging, vast and incorporates several features which closely resembles human behavior to identify quality and appropriate manuscripts. Initial set of experiments on curated and scholarly data show promising results. We are sure that there more issues to address, many scope of improvements which would eventually lead us one step closer to this ambitious vision: *to cut through the clutter of bad literature and accelerate scientific discovery* and eventually bring AI more close to the academic peer review system.

KEYWORDS

computational support, academic peer review, novelty detection, scope detection, quality prediction, research articles

1 INTRODUCTION

Peer review is the backbone of scholarly publishing and despite criticisms, is still the only widely accepted method for research validation. However peer review has its own shortcomings and the most prominent problems are: a seemingly time consuming process, inherent human bias, exponential rise of paper submissions, etc. With the wide spread applications of Artificial Intelligence across every aspects of life, we are intrigued to think *how the AI techniques could assist the present day peer review system?* Our idea is not to try and develop an automated system for peer review (which is by far not desirable as well as questionable), but to find out the areas where AI could act as an aid with reasonable certainty. The very first stage in peer review is the initial screening usually performed by editors. Day by day the number of submissions made to each journal is rising. Editors, who are usually full-time academicians, are overwhelmed with this huge number of manuscripts they had to go manually through. With the ever expanding volume of research articles it is increasingly becoming difficult for the editors to keep pace with the latest research and trends. These also hinders editorial response in a reasonable time frame. At this stage journal editors are entrusted to take either of the two decisions: whether to forward the manuscript to expert reviewers for meticulous evaluation or to outright reject the paper from the desk. Analysis of about 5.5k desk rejected articles along with corresponding author-editor-reviewer interactions across 11 different Computer Science journals led us to believe that there exists at least five common reasons that leads to Desk Rejection of academic manuscripts.

(1) **Quality**/Standard and impact of the article under review.
(2) Appropriateness of the article to the journal being sent (Aim and **Scope**).
(3) Percentage overlap with existing articles (Plagiarism).
(4) Spelling, grammar and language of the article under review.
(5) Visually discriminative features of the article such as template mismatch (article not being prepared according to journal guidelines and formatting requirements), articles not having the standard components of a proper scientific communication.

While (4) is somewhat a supplementary reason, but others are very intense for desk rejection. Once a manuscript successfully goes past the editors desk, one factor that stands tall above anything else is **Novelty**. As for (3), (4) and (5) we already have *state-of-the-art* systems in deployment, we begin our investigation with the three crucial factors *Quality*, *Scope*, and *Novelty* on scholarly data. The objective of this doctoral work is to investigate inclusion of artificial intelligence as an assistant to the editors, reviewers and authors so as to reduce the shortcomings associated and bring more transparency to the academic peer review system. A system of this kind would aid the editors to:

- identify out-of-scope submissions
- identify submissions with visible low quality content
- identify potential related works or influential works with respect to a submission

We embark upon to detect *document-level novelty* from text documents (mostly objective newspaper articles). Knowledge gained from such experiments on document-novelty may be further leveraged towards the detection of novelty from scholarly articles. We begin with developing a benchmark resource for document-level novelty detection from news articles and design appropriate machine learning solutions to classify a document as *novel.*

2 PROPOSED RESEARCH

We begin our work with the following Research Goals (RG). However due to the complexity of the task for research articles and lack of available datasets, we proceed to explore document-level novelty detection in case of news articles.

2.1 RG 1 : Document-Level Novelty Detection

We view Document-Level Novelty Detection as a classification task of automatically labeling a target document as novel or non-novel. The decision is not universal and should always be with respect to a set of source documents already seen by the system. The investigation here is: *What makes a document appear novel to a reader? What is that amount of new information in a novel document which distinguishes it from a non-novel or partially novel one?* More specifically we are seeking a *Novel Document Classification* task.

Detecting *novelty* of a research article is something very non-trivial and could not be reached via merely text mining. We need to explore deeper semantic and pragmatic knowledge engineering framework for that. Also there are no available datasets that addresses novelty detection in research articles. So to begin with, we explore available corpora and come across the datasets of TREC sentence-level novelty detection [51] task and RTE-TAC sentence level novelty sub tracks [4]. These tasks were initiated from an information retrieval perspective of selecting the new sentences out of a collection. But our goal is somewhat different. To explore: *Whether a document could be called novel or non-novel?* with respect to a set of documents already seen (aka the knowledge base). The decision should consider semantics of the texts involved (both source and target); not just the lexical surface form.

Analysis and observation of data lead us to believe that there exists at least four properties that characterizes novelty detection from texts. *Relevance, Diversity, Relativity and Temporality*. The target document should be relevant to the context or source documents. For example, seeking novelty between two documents, one talking about *jaguar*, the animal and the other about *jaguar*, the car is futile as one is not relevant to the other. Quite obvious that each one would contain different information than the other. *Diversity* directly correlates with the new information content. *How new or diverse is the target document in terms of information content?* Also the amount of new information content is important while deciding the novelty of an entire document (*Relativity*). *Is the amount of new information sufficient to call the document novel?* Novel information is usually a *temporal* update over existing relevant knowledge. So we call a document as novel when it has sufficient diverse yet relevant information. With these views we set out to discover the state of novelty of a document against a predefined set of source documents. Our intention is to make the machine learn this perspective of ours towards document-novelty. We also create a resource for *document-level novelty detection* and discuss the same in Section . The knowledge gained from working with news article data could be extended for research articles with appropriate assumptions. Semantic-level plagiarism detection comes close to our investigation objective.

Please note that the current investigation is to address document-level novelty detection for objective newspaper texts. Since there is no available dataset for detecting research novelty and that the problem in the research domain is non-trivial in nature, we intend to explore it in future. The same problem would require deeper investigation perspective encompassing cognition, derivational logic, knowledge discovery blended approach.

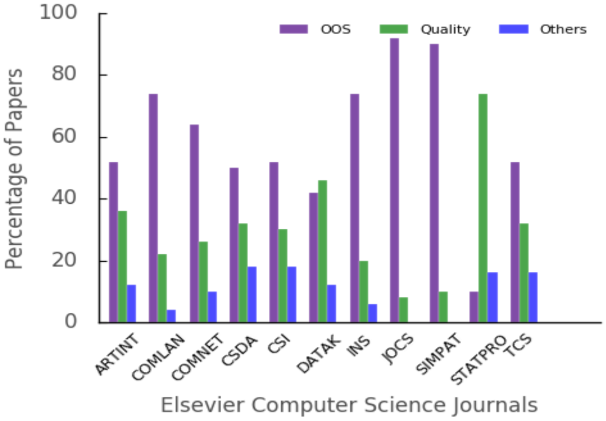

Figure 1: Percentage of desk rejected papers that were rejected due to "out-of-scope" (OOS), "quality" (Q), and "other" for 11 Elsevier computer science journals: Artificial Intelligence (ARTINT), Computer Languages (COMLAN), Computer Networks (COMNET), Computational Statistics and Data Analysis (CSDA), Computer Standards and Interfaces (CSI), Data and Knowledge Engineering (DATAK), Information Sciences (INS), Journal of Computational Sciences (JOCS), Simulation Modeling Practice and Theory (SIMPAT), Statistics and Probability Letters (STATPRO), and Theoretical Computer Science (TCS)

2.2 RG 2: Scope Detection

With this problem we set out to measure the *appropriateness of a manuscript to a journal*. The question is : *Whether a submission is a good fit to the topic coverage and aspirations of the intended journal?*Our study on a sample set of Computer Science journals show that more than 50% of desk rejections (see Figure 1) owes to articles being not within the scope of the journal to which were originally submitted. *Out-of-scope* means the manuscript does not fall into the domain of interest of the journal or could not cater to the interest of the associated community. Here too we view the problem as a classification of an article into *In-Scope* and *Out-Scope* classes. Our investigation takes into account the desk rejected articles and accepted articles of a given journal where the content of accepted articles serve as the benchmark of reference. Our venture is to extract meaningful features from almost every section of a manuscript that could contribute to determine its scope. This way we believe we could develop a system which could stand out as an assistant to both editors and authors for taking wise decisions regarding the appropriateness of a manuscript to a journal.

2.3 RG 3: Quality Prediction

The next problem that we tackle in this work is predicting the quality of a manuscript. While analyzing the desk rejected articles we see that many articles are rejected because they are deemed not good enough to the standards of the journal. Even we come across such articles that do not follow the minimum standards or components of a proper scientific communication. This in no way overlaps with the novelty criterion. Bibliography, context citations, author credibility, language and grammar, etc. plays a major role here. Inspired from the pioneering works [3, 17, 19, 20, 48, 53, 60] on impact prediction of scientific articles, we seek to investigate the following for *Quality* :

- **Quality of Bibliography** : Whether relevant and important citations are present or not? The quality of the citations. The temporal distance of the citations.
- **Identifying Impact and Incidental Citations** : Whether the citations are relevant to the context or incidental [54].
- **Argumentation Mining** : Extracting argumentative structures from the scientific discourse and verifying their relevance and validity of claims w.r.t the context [28].
- **Academic Hedging** : Identifying effect of hedges in scientific writings [29]. Often authors are not confident about the claims they make and they take help of "vague language" to justify their propositions.
- Effect of author credibility on scientific discourse quality.

Based on these quality features we intend to measure the *depth* of a scientific article. While there are more factors to consider we plan to begin our investigations with these.

2.4 RG 4 : Resource Creation

Another goal of this work is to create benchmark resource for document-level novelty detection. Although benchmark data for novel sentence retrieval are released by TREC and RTE-TAC, but they are developed from an IR perspective and do not cater to our document level investigation needs. The only publicly available dataset for document-level novelty detection is APWSJ [58]. However this dataset too was developed for IR experiments and is skewed towards novel documents (only 9.07% non-novel documents). Hence we felt the need to design a dataset for document-level novelty detection from news articles encompassing the aspects as discussed in Section 2.1. We discuss our resource in Section 5.4.

3 BACKGROUND AND RELATED WORK

Artificial Intelligence techniques in academic peer review system in entirety is a less explored frontier. Although we find that there has been quite a good amount of work on the several components we identify in our investigation. Sentence-level novelty detection has been a popular problem in information retrieval. However we find that very little has been explored on the problem at document-level which is central to this research on scholarly data. Similarly although there are some excellent works on scholarly venue recommendation systems [9, 38, 57] for academic manuscripts, we find scope detection is not explored as a classification problem to weed out *out-of-scope* submissions. We stand on the *shoulder of these giants* and proceed with our investigation.

3.1 Novelty Detection

Research in novelty mining could be traced back to the Topic Detection and Tracking (TDT) [52] evaluation campaigns where the concern was to detect new event or First Story Detection (FSD) with respect to online news streams. Techniques mostly involved grouping the news stories into clusters and then measuring the belongingness of an incoming story to any of the clusters based on some preset similarity threshold. Some notable contributions from TDT are [1, 2, 55, 56].

The task gained prominence in the novelty tracks of Text Retrieval Conferences (TREC) from 2002 to 2004 [11, 44, 50, 51] although the focus was sentence-level novelty detection. The goal of these tracks was to highlight the relevant sentences that contain novel information, given a topic and an ordered list of relevant documents. Next came the novelty sub tracks of Recognizing Textual Entailment-Text Analytics Conference (RTE-TAC) 6 and 7 [4] where Textual Entailment was viewed as one close neighbor to sentence level novelty detection.

At the document level an interesting work was carried out by [56] via topical classification of on-line document streams and then detecting novelty of documents in each topic exploiting the named entities. Another work by [58] viewed novelty as an opposite characteristic to redundancy and proposed a set of five redundancy measures ranging from the set difference, geometric mean, distributional similarity in order to calculate the novelty of an incoming document with respect to a set of memorized documents. They also presented the first publicly available Associated Press-Wall Street Journal (APWSJ) news dataset for document level novelty detection. [47] applied a document to sentence level framework to calculate the novelty of each sentence of a document which aggregates to detect novelty of the entire document. [33] computed novelty score based on the inverse document frequency scoring function. Another work by [49] presents a comparison study of different novelty detection methods evaluated on news articles where language model based methods perform better than the cosine similarity based ones. More recently [16] conducted experiments with information entropy measure to calculate innovativeness of a document. Each of these works evaluated their methods on separate datasets and to the best of our knowledge none except APWSJ are publicly available.

Novelty detection is also studied in works related to diversity in information retrieval literature. Idea is to retrieve relevant yet diverse documents in response to user query. The work on Maximal Marginal Relevance [7] was the first to explore diversity and relevance for novelty. Some other notable works along this line are [8, 12, 13]. Zhao and Lee [59] recently proposed an intriguing idea of assessing the novelty apetite of an user based on a curiosity distribution function derived from curiosity arousal theory and Wundt curve in psychology research. The work that we present here significantly differs from the existing literature as along with hand crafted features, we provide a deep neural network solution to the problem which learns the notion of novelty and non-novelty from the data itself.

3.2 Scope Detection

Most of the reputed journal publishers have their own systems that suggest relevant journals to an author against her work. Examples could be given of Journal Finder by Elsevier[1], Springer Journal Suggester[2], EDANZ Journal Selector[3],etc. Also some web-services like JANE (Journal/Author Name Estimator)[4] [43], eTBLAST [23], GoPubMed [18], HubMed [22], Pubfinder [27], etc. suggest relevant

[1] http://journalfinder.elsevier.com/

[2] http://journalsuggester.springer.com/

[3] https://www.edanzediting.com/journal-selector

[4] http://jane.biosemantics.org/

biomedical literatures from PubMed[5] or MEDLINE[6] databases upon user query (typically the title and abstract of the article for which the user wants to find a 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.

3.3 Quality Prediction

We are inspired from the works on scientific impact prediction and argumentation mining from scientific articles which we deem relevant to our investigation for quality prediction of academic manuscripts. We take inspiration from these [3, 17, 19, 20, 48, 53, 60] exceptional works on citation analysis and citation networks. Argumentation mining is another avenue which we would like to investigate to judge the viability of claims presented in a scientific discourse. We seek to identify and focus on the analysis of argumentation structures in scientific publications on a fine-grained level. The goal is to reveal how an author connects her thoughts in order to create a convincing line of argumentation. Such a fine-grained analysis of the argumentation structure will enable new ways information access, and could be integrated, for example, in summarization or faceted search applications as part of digital libraries. Some notable works we look forward to here are [28, 35, 45, 46].

4 DATA

The dataset we create for document-level novelty detection (RG1) is described in Section 5.4. Some other datasets that we use for our novelty experiments are the paraphrase detection Webis-Crowd Paraphrase Corpus [6] and the APWSJ [58].

For RG2 and RG3 we are thankful to Elsevier for providing us the requisite data. We consider all the accepted articles published till 2017 from 11 different Elsevier Computer Science journals which amounts to about 60K full text articles. For each of these journals we consider 1000 rejected articles along with their author-editor-reviewer interactions. We actually went through about 7000 review reports (author-editor-reviewer interactions) of accepted, rejected and rejected-after-review papers consisting of more than 2.1 million lines of review data to investigate the generic causes of refutation. These interactions form the primary source of our feature design. We transform the articles, originally in .pdf form to .json and .xml format for information extraction. To investigate RG2 and RG3 we create various dictionaries, lists, structures, temporal repositories for experiments on each journal data. We also develop a *word2vec*[39] model trained on all accepted articles of 223 Elsevier Computer Science journals covering more than 7 million full text articles with more than 6.5 million vocabulary and use in our experiments. We also went through about 7000 review reports (author-editor-reviewer interactions) of accepted, NFWD and declined-after-review papers consisting of more than 2.1 million lines of review data to investigate the generic causes of refutation and then develop a good set of features.

We also plan to use the following open source datasets for our experiments with scholarly information extraction and mining : CORE [36] and the Open Corpus by Semantic Scholar[7].

5 RESEARCH METHODOLOGY

5.1 Novelty Detection

As stated earlier we consider document level novelty detection as a classification task. We employ both feature based techniques and deep learning methods to label a document as novel or non-novel. One Natural Language Processing (NLP) task that closely relates to our problem of semantic level novelty detection is *Textual Entailment* [15]. The high level semantic interactions required to determine whether a text is entailed from other is a close approximation to determine whether a given text is *non-novel* or *novel*. This relationship is true when we experiment with straightforward objective texts (for e.g., newspaper article texts). But when we consider intelligent texts having a huge premise like research articles, straightforward techniques to recognize textual entailment would not be sufficient enough. Detecting premise of a scholarly text from the huge volume of scientific knowledge is a challenging task in itself [31]. We firmly believe that straightforward text mining techniques would be inadequate to address this problem. We ned constructs like Knowledge Graph [35] built upon a huge volume of scientific articles to deduce conclusions from the inter relationships of scientific entities. We gradually intend to explore these complex tasks of Scientific Entailment Prediction for novelty detection with the recently released SciTail dataset [10] and Scientific Premise Selection with the Mizar corpus [30]. However to begin with, we try to make best use of the available resources for novelty detection and frame it as a classification task.

5.1.1 Feature based techniques (Approach-I).

We view novelty as an opposite characteristic to semantic textual similarity. We curate several features from a target document (with respect to predefined set of source documents) like paragraph vector (*doc2vec*) similarity, KL divergence, summarization similarity (concept centrality), lexical n-gram similarity, new words count, etc and build our classifier based on Random Forests. Our method yields promising results w.r.t baselines assumed. The details of the feature descriptions could be found in Ghosal et al. [25]

5.1.2 Deep Learning Approaches.

Here instead of manually extracting features from the target document with respect to the source document(s) we explore deep learning techniques to automatically extract features from the data. We experiment with two approaches and represent our target documents as semantic vectors. We train our sentence encodings on the semantically rich, large scale (570k sentence pairs) Stanford Natural Language Inference (SNLI) dataset [5]. As discussed earlier, Natural Language Inference or Textual Entailment is one such task that involves high level semantic and pragmatic knowledge which supposedly captures the complex semantic interactions necessary for determining semantic redundancy or non-novelty. We generate sentence encodings by feeding Glove [42] word vectors to a Bi-Directional LSTM followed by max pooling [14].

Approach -II (RDV-CNN) : We arrive to a certain document

[5] https://www.ncbi.nlm.nih.gov/pubmed/

[6] https://www.nlm.nih.gov/bsd/pmresources.html

[7] http://labs.semanticscholar.org/corpus/

level semantic representation (inspired from [40]) that models both source and target information in a single entity which we coin as the *Relative Document Vector (RDV)*. Each sentence in the target document is represented as :

$$RSV_k = [a_k, b_{ij}, |a_k - b_{ij}|, a_k * b_{ij}]$$

where RSV_k is the Relative Sentence Vector of sentence k in the target document. a_k is the sentence embedding of the target sentence k and b_{ij} is the sentence embedding of the i-th sentence in source document j. The selection of premise source sentence ij is done via the highest cosine similarity. We stack the Relative Sentence Vectors (RSV) corresponding to all sentences in a target document to form the RDV. The RDV becomes the input to a deep Convolutional Neural Network (CNN) [34] for automatic feature extraction and subsequent classification of a document as *novel* or *non-novel*. We achieve encouraging results on APWSJ corpus as listed in Table 1. On high level paraphrase detection task our approach supersedes the other approaches (Table 2). The results clearly indicates that our deep network RDV-CNN is able to learn the semantic interactions necessary to comprehend novelty at the document level. On TAP-DLND 1.0 we report the results in Table 3. Please refer to Ghosal et al. [24] for a detailed descrition of this work.

Measure	Recall	Precision	Mistake
Set Difference	0.52	0.44	43.5%
Cosine Distance	0.62	0.63	28.1%
LM:Shrinkage	0.80	0.45	44.3%
LM:Dirichlet Prior	0.76	0.47	42.4%
LM:Mixture Model	0.56	0.67	27.4%
RDV-CNN	**0.58**	**0.76**	**22.9%**

Table 1: Results for Redundant class on APWSJ, LM **→ Language Model,** $Mistake$ **→100-Accuracy. Except for RDV-CNN, all other numbers are taken from [58]**

Systems	P	R	F_1
Set Difference + LR [58]	0.71	0.52	0.60
Geometric Distance + LR [58]	0.69	0.75	0.72
LM: Dirichlet Prior + LR [58]	0.74	0.77	0.75
Novelty (IDF) + LR [33]	0.65	0.55	0.59
Paragraph Vector+LR (Baseline) [37]	0.59	0. 75	0.72
RDV-CNN	**0.75**	**0.84**	**0.79**

Table 2: Results for Paraphrase class on Webis-CPC (in %), IDF **→ Inverse Document Frequency,** LR **→ Logistic Regression**

Approach-III: Here we experiment with another deep neural approach based on attention mechanism inspired from [41]. For an actually redundant document, we contend that neural attention mechanism would be able to identify the sentences in source documents that has identical information and is responsible for *non-novelty* of the target document. Our contention is that alignment via attention of a target document texts with potential source document(s) would facilitate the creation of a joint source encapsulated semantic representation of the target document that would enable a neural network to learn patterns of novelty and redundancy in document(s). This approach is very simple with an order of fewer parameters as compared to other complex deep neural architectures for modeling natural language inference tasks. It relies on only learning of sentence alignments, inspired from works on attention in Machine Translation literature. As a baseline we take: joint encoding of source and target sentences fed to a BiLSTM network. The output of the last layer is then fed to feed forward network followed by classification via softmax. Approach-III supersedes our other two methods on TAP-DLND 1.0 by a good margin. Also all our methods appear promising across all datasets we choose for document level novelty detection. These works on objective newspaper texts is for gaining requisite insights on novelty before we proceed to mine scientific literature for the same task.

Methods	P	R	A
Baseline	0.81	0.82	81.4
Entropy based method[16]	0.67	0.67	68.2
Approach-I [25]	0.79	0.79	79.2
Approach-II (RDV-CNN)	**0.85**	**0.85**	**84.5**
Approach-III	**0.87**	**0.87**	**87.4**

Table 3: 10-*fold* cross validation results on TAP-DLND 1.0, P **→ Average Precision,** R **→Average Recall,** A **→Accuracy(%)**

5.2 Scope Detection

We extract features from almost every section of a scientific manuscript that could contribute to identify its domain: *Author, Content and Bibliography*. These features will pave the way for a better venue recommendation system for both the editors and the authors. However till now we are able to report the results for the classification experiments. Our point of departure for this particular work was the bibliography section of research articles. Articles already published by a journal signify that they are within-the-scope of that journal. Our *out-of-scope* data are those desk-rejected manuscripts for which the editor(s) felt are not a good fit to the topic-coverage and aspirations of the journal and hence circumvented their progress further in the review process. We hypothesize that with obvious exceptions *if an article belongs to a particular domain then majority of its references would fall in that certain domain.* Coupled with other factors, our approach *ScopeJr* achieves *state-of-the-art* performance across six different journals. We extract several features from different sections of a manuscript with respect to the stored history information of accepted articles of the particular journal. Please refer to [26] for details of the feature definitions. The Scope features are :

(1) Number of keywords match weighted by frequency across all accepted articles(author-listed+extracted) (*wt_kw_m*)

(2) Semantic distance of a candidate article from the cluster of history accepted articles (*clust_dist*)

(3) Overlap of bibliographic paper titles with that of history articles (*bib_tit_sc*)

(4) Overlap of bibliographic venues (journals/conferences) with that of history articles (*bib_jr_sc/bib_conf_sc*)

(5) Average of the number of times the authors published in the particular journal (author domain publication frequency)(*adpf*)

We design a novel function **Citation Effect** which counts the frequency of citations within the body of the paper and employ it as a weight of (4) and (5). *In-domain* keywords, referred titles and venues have higher occurrence across all the accepted articles. Table 4 displays the performance of our approach against the Elsevier journal recommendation system. Our approach *ScopeJr* is based on Random Forest classifier. For each of the journals we take 1000 accepted papers as *in-scope* data and 1000 *out-of-scope* articles actually rejected from the desk. We extract features and perform the experiments in a *10-fold cross-validation* classification set up. Finally we compare the classification performance of our proposed system with the *state-of-the-art* **Elsevier Journal Finder (EJF)**[32] on the same dataset and report the results (Table 1). EJF is a *state-of-the-art* recommender system provided by Elsevier solutions to the academic fraternity that recommends highly relevant journals to the authors for their papers. Elsevier Journal Finder takes as input the *Title* and *Abstract* of a prospective scientific article (Y) and presents a list of 10 relevant Elsevier journals (J) to the user as output which s/he may consider for submitting her/his article. Although the recommended journals are limited only to Elsevier published ones, but it is to be noted that Elsevier has more than 2900 peer-reviewed journals that cover almost all major scientific domains. Although we had *true class* labels from Elsevier data, we follow heuristics to determine the **EJF** *predicted* class label of a prospective article Y : *If EJF suggests J for Y* $\longrightarrow$*Y is* ***In-Scope*** *of J otherwise,* ***EJF*** *deems Y to be* ***Out-of-Scope*** *for J* Thorough analysis of data and experimental results led us to the following observations:

Journals	Methods	P(OS)	R(OS)	Acc.(%)
ARTINT	Elsevier Journal Finder	0.542	0.621	63.64
	ScopeJr	0.885	0.856	† 87.25
COMNET	Elsevier Journal Finder	0.341	0.431	44.43
	ScopeJr	0.823	0.803	† 81.49
STATPRO	Elsevier Journal Finder	0.433	0.527	53.56
	ScopeJr	0.837	0.843	† 83.93
TCS	Elsevier Journal Finder	0.556	0.648	66.82
	ScopeJr	0.869	0.876	† 87.20
CSI	Elsevier Journal Finder	0.512	0.674	65.64
	ScopeJr	0.815	0.951	† 86.75
SIMPAT	Elsevier Journal Finder	0.532	0.656	64.86
	ScopeJr	0.726	0.767	† 72.23

Table 4: Scope-Check figures for *out-of-scope* (OS) class across 6 journals, $P \rightarrow Precision$, $R \rightarrow Recall$. The Accuracy values (†) for *ScopeJr* are statistically significant over EJF performance (two-tailed t-test, p<0.05)

(1) *Bibliographic* features have induced significant improvements (Figure 3) due to the fact that *Bibliographic feature* values were deduced from within the body section of the scientific articles. *When a certain portion of a scientific article cites a reference, the scope of that portion is influenced by the domain of that reference article. The domain of the cited reference exerts local influence on that portion of the scientific article.* So if many in-domain references are cited in distributed portions of a research article, quite possibly the entire research article falls in the same domain. We measure *in-domain* or *in-scope* by simply counting occurrences of features (Section 2.2) across a certain journal, higher the better.

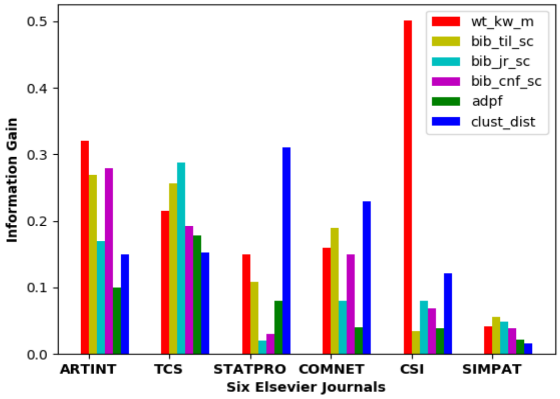

Figure 2: Significance of features observed by ranking features based on Information Gain

(2) For all the journals our approach outperforms the **EJF** in terms of precision, recall and accuracy values. This could be attributed to the fact that EJF only considers the *Title* and *Abstract* portions of a research article and uses the **Elsevier Finger Print Engine**[8] based on identification of *Noun Phrases* from those sections. Our method goes beyond this and uses *Bibliographic, Author and Content* information which definitely contributes to categorization of an article in a particular domain.
(3) Some journal specific features (like presence of mathematical expressions for STATPRO) may further improvise.
(4) For journals having very wider scope (for e.g., Computer Science Review or Nature or Science) or multi-disciplinary in nature, this approach may not be fruitful.
(5) Scope of a journal gets more compact and streamlined with time. Hence experimenting with only recent articles instead of historical ones may boost the performance.
(6) Journals SIMPAT and CSI accept papers across many domains. Hence their *Bibliography* sections are distributed into different domains. However for ARTINT, STATPRO and COMNET we find *Bibliography* generates a restricted *domain-specific* set and hence proves more effective.

5.3 Quality Prediction

The quality prediction task is coupled with the scope detection task as quality of a paper stands out to be another deciding factor for desk rejection of scientific manuscripts. We probe into *Author, Affiliation, Content and Bibliography* features based on paper meta data. *The better is the bibliography section of a paper, better are its chances to escape desk rejection. Some content characteristics and location of impact/in-domain citations exerts a "local influence" on portions of the paper that greatly determines its credibility as well as scope.* By "better" we mean high impact as well as recent citations. We also devise a way to identify citations which are actually important to a paper and citations which are just mentioned incidentally. We

[8] https://www.elsevier.com/solutions/elsevier-fingerprint-engine

design features based on author and affiliation credibility, bibliography credibility and content. We see that augmentation of quality features with scope greatly increases the probability of identifying rejected submissions by our system [26]. Inspired from the works on scientific impact prediction [21, 54], we extract meta information from the author profile, bibliography section and content to arrive to a set of features that to some extent plays a role in determining the quality of the prospective manuscript. **Author** credibility features are *Primary author's h-index, citation count. Average, Maximum* of *h*-indices, citation counts of all the authors. Similarly for the **Affiliation** credibility, we take the *research scores* and the *number of research articles produced* by the author-affiliated institutions from the Times Higher Education World University Rankings and Scopus[9] respectively (Primary author, maximum, average). To measure the **bibliography quality** of the manuscript we take the citation counts of the referred articles, reputation of the venues where published (CORE Computer Science rankings, h-index and SCImago Journal Rank values[10]) and temporal distance of the references from the submission date. **Content** features that we extract include the number of uncited references, mathematical equations, figures and tables. Please refer to [26] for details of the feature definitions. Due to paucity of space we restrain to define the features in detail.

5.4 Resource Creation

As we mention earlier, detecting novelty for research is non-trivial. As because novelty in research is not just textual but conceptual. We may need a joint approach of cognition, derivational logic, knowledge discovery to identify novelty of scientific claims. However, in this particular work we focus on detecting document level textual novelty from objective newspaper texts.

We create a benchmark resource for Document Level Novelty Detection (DLND) [25] and call it TAP-DLND 1.0 (TAP : Tirthankar-Asif-Pushpak, the primary investigators). The dataset is balanced and consists of 2736 novel documents and 2704 non-novel documents. For each novel/non-novel document there are three source documents against which the target documents are annotated. The state of *novelty* for each target document is to be measured against those source documents i.e. once the system has already seen the designated source documents for a particular event, it is to judge whether an incoming on-topic document is novel or not. The structure of TAP-DLND 1.0 is in Figure 3. We crawl news events from 10 different categories and annotate target documents with respect to the designated source documents. We leave out partially novel or ambiguous cases from our annotations.

6 RESEARCH CONTRIBUTIONS

(1) Algorithms/methods for document-level novelty detection.
(2) Studies on possible reasons of rejections on a massive data and AI ways to assist the editors/reviewers. Investigation on AI techniques to streamline the various aspects of academic peer review system.

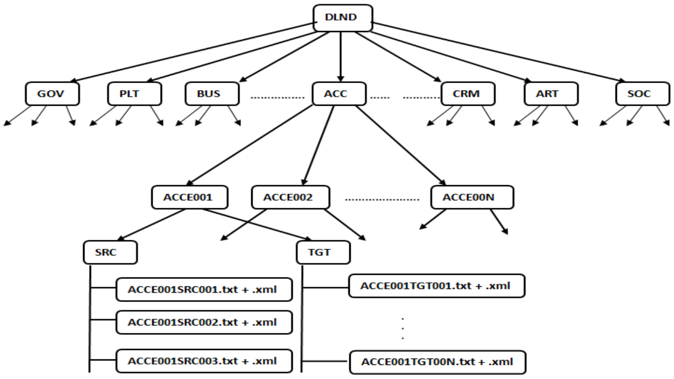

Figure 3: TAP-DLND 1.0 corpus structure

(3) Resources for document-level novelty detection
(4) Novel methods for determining whether a manuscript is within the scope of a journal. A robust journal recommender system for academic manuscripts with the identified features
(5) Methods for assessing the quality of a scientific manuscript based on bibliographic, content, arguments and author profile features.

7 FUTURE PLANS

We are currently exploring and intend to investigate the following:

- Deducing a more effective target document vector representation with respect to source information encompassing semantic aspects
- Argumentation Mining, Academic Hedging and Text Simplification for quality prediction of research articles
- Extending the scope factors for a more accurate venue recommendation system
- Exploring Knowledge Graphs for novelty detection.
- Preparing a sentence level annotated corpus for document level novelty detection encompassing all the four aspects discussed in Section 5.4
- An end-to-end deep neural architecture to compute the novelty score of a document based on texts already seen by the system.
- Extending the novelty detection architecture to select relevant source documents to reduce the computational overhead when the number of source is very large (Relevance Judgment). Deducing a mechanism to select and encode relevant information from multiple premises for a target text (Multi premise entailment relationships).
- A multimodal (image+text) and multiview multiobjective investigation into scope detection of research articles.

8 TIMELINE

Table 5 estimates timeline of this proposed research.

[9] https://www.timeshighereducation.com/world-university-rankings/2018, https://www.elsevier.com/solutions/scopus
[10] http://www.scimagojr.com/, http://portal.core.edu.au/conf-ranks/

Month/Year	Investigation and Tasks
2016	Literature Survey, Problem Identification, Dataset Creation, Course Work
January 2017-June 2017	Novelty Detection (Feature-Based+Deep Neural Method)
July 2017 - December 2017	Scope Detection and Quality Prediction
January 2018-June 2018	Attention-based Novelty Detection, Novelty Scoring of Documents
July 2018 - February 2019	Multimodality and Multiview Multiobjective Scope Detection
March 2019-August 2019	Knowledge Graphs for Novelty Detection
September 2019-December 2019	AI in Peer Review (Predict Accept/Reject probability and Aspect Scores)
January 2020-June 2020	Argumentation Mining, Academic Hedging, Citation Analysis, Text Simplification for Quality Prediction
July 2020 - December 2020	Writing Thesis and Complete Pending works

Table 5: Proposed Timeline of Doctoral Research

9 CONCLUSIONS

This work is an intersection of various challenging aspects where AI could play a part in peer review. Extending computational support for detecting novelty, scope and quality of a research article would claim more transparency in the peer review ecosystem. Our proposed methods show promise to deliver. We look forward to expert community evaluation of our take on this problem.

ACKNOWLEDGMENTS

The author is supported by Visvesvaraya PhD fellowship, Ministry of Electronics and Information Technology (MeitY), Government of India (GoI). We are grateful to Elsevier for providing us the required data for conducting our experiments. I am grateful to my supervisors Dr. Asif Ekbal, Prof. Pushpak Bhattacharyya and Dr. Sriparna Saha for their constant guidance and motivation. I would like to thank our annotators Ms. Amitra Salam and Ms. Swati Tiwari for their efforts to construct the TAP-DLND 1.0 dataset. I also extend my gratitude to ACM SIGIR for awarding me the Student Support Travel Grant to attend and present my work in JCDL 2018.

REFERENCES

[1] James Allan, Victor Lavrenko, and Hubert Jin. 2000. First story detection in TDT is hard. In *Proceedings of the ninth international conference on Information and knowledge management*. ACM, 374–381.

[2] James Allan, Ron Papka, and Victor Lavrenko. 1998. On-line new event detection and tracking. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 37–45.

[3] Juyoung An, Namhee Kim, Min-Yen Kan, Muthu Kumar Chandrasekaran, and Min Song. 2017. Exploring characteristics of highly cited authors according to citation location and content. *JASIST* 68, 8 (2017), 1975–1988. https://doi.org/10.1002/asi.23834

[4] Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge.. In *TAC*.

[5] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. 632–642. http://aclweb.org/anthology/D/D15/D15-1075.pdf

[6] Steven Burrows, Martin Potthast, and Benno Stein. 2013. Paraphrase acquisition via crowdsourcing and machine learning. *ACM Transactions on Intelligent Systems and Technology (TIST)* 4, 3 (2013), 43.

[7] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 335–336.

[8] Praveen Chandar and Ben Carterette. 2013. Preference based evaluation measures for novelty and diversity. In *SIGIR*.

[9] 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.

[10] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457* (2018).

[11] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track.. In *TREC*, Vol. 4. 74.

[12] Charles L. A. Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. 2011. A comparative analysis of cascade measures for novelty and diversity. In *WSDM*.

[13] Charles L. A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *SIGIR*.

[14] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 670–680. https://aclanthology.info/papers/D17-1070/d17-1070

[15] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. Springer, 177–190.

[16] Tirthankar Dasgupta and Lipika Dey. 2016. Automatic Scoring for Innovativeness of Textual Ideas. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

[17] Ying Ding, Erjia Yan, Arthur R. Frazho, and James Caverlee. 2010. PageRank for ranking authors in co-citation networks. *CoRR* abs/1012.4872 (2010). arXiv:1012.4872 http://arxiv.org/abs/1012.4872

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

[19] Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. 2015. Will This Paper Increase Your *h*-index?: Scientific Impact Prediction. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2-6, 2015*. 149–158. https://doi.org/10.1145/2684822.2685314

[20] Yuxiao Dong, Reid A. Johnson, and Nitesh V. Chawla. 2016. Can Scientific Impact Be Predicted? *IEEE Trans. Big Data* 2, 1 (2016), 18–30. https://doi.org/10.1109/TBDATA.2016.2521657

[21] Yuxiao Dong, Reid A Johnson, and Nitesh V Chawla. 2016. Can scientific impact be predicted? *IEEE Transactions on Big Data* 2, 1 (2016), 18–30.

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

[23] 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.

[24] Tirthankar Ghosal, Vignesh Edithal, Asif Ekbal, Pushpak Bhattacharyya, George Tsatsaronis, and Srinivasa Satya Sameer Kumar Chivukula. 2018. Novelty Goes Deep. A Deep Neural Solution To Document Level Novelty Detection. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 2802–2813. https://aclanthology.info/papers/C18-1237/c18-1237

[25] Tirthankar Ghosal, Amitra Salam, Swati Tiwari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. TAP-DLND 1.0: A Corpus for Document Level Novelty Detection. *arXiv preprint arXiv:1802.06950. To appear in the Proceedings of the 11th International Language Resources and Evaluation Conference (LREC 2018)* (2018).

[26] Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, Srinivasa Satya Sameer Kumar Chivukula, Georgios Tsatsaronis, Pascal Coupet, and Michelle Gregory. 2018. Can your paper evade the editors axe? Towards an AI assisted peer review system. *arXiv preprint arXiv:1802.01403*

(2018).

[27] 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.

[28] Nancy Green. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*. 11–18.

[29] Ken Hyland. 1998. *Hedging in scientific research articles*. Vol. 54. John Benjamins Publishing.

[30] Mihnea Iancu, Michael Kohlhase, Florian Rabe, and Josef Urban. 2013. The Mizar mathematical library in OMDoc: translation and applications. *Journal of Automated Reasoning* 50, 2 (2013), 191–202.

[31] Geoffrey Irving, Christian Szegedy, Alexander A Alemi, Niklas Een, Francois Chollet, and Josef Urban. 2016. Deepmath-deep sequence models for premise selection. In *Advances in Neural Information Processing Systems*. 2235–2243.

[32] Ning Kang, Marius A Doornenbal, and Robert JA Schijvenaars. 2015. Elsevier journal finder: recommending journals for your paper. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 261–264.

[33] Margarita Karkali, François Rousseau, Alexandros Ntoulas, and Michalis Vazirgiannis. 2013. Efficient Online Novelty Detection in News Streams.. In *WISE (1)*. 57–71.

[34] Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[35] Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2015. Linking the thoughts: Analysis of argumentation structures in scientific publications. In *Proceedings of the 2nd Workshop on Argumentation Mining*. 1–11.

[36] Petr Knoth and Zdenek Zdrahal. 2012. CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine* 18, 11/12 (nov 2012). https://doi.org/10.1045/november2012-knoth

[37] Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.. In *ICML*, Vol. 14. 1188–1196.

[38] Eric Medvet, Alberto Bartoli, and Giulio Piccinin. 2014. Publication venue recommendation based on paper abstract. In *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*. IEEE, 1004–1010.

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[40] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin. 2015. Natural language inference by tree-based convolution and heuristic matching. *arXiv preprint arXiv:1512.08422* (2015).

[41] Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. 2249–2255. http://aclweb.org/anthology/D/D16/D16-1244.pdf

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

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

[44] Ian Soboroff and Donna Harman. 2005. Novelty detection: the trec experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 105–112.

[45] Christian Stab, Christian Kirschner, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation Mining in Persuasive Essays and Scientific Articles from the Discourse Structure Perspective.. In *ArgNLP*. 21–25.

[46] Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics* 28, 4 (2002), 409–445.

[47] Flora S Tsai and Yi Zhang. 2011. D2S: Document-to-sentence framework for novelty detection. *Knowledge and information systems* 29, 2 (2011), 419–433.

[48] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying Meaningful Citations. In *Scholarly Big Data: AI Perspectives, Challenges, and Ideas, Papers from the 2015 AAAI Workshop, Austin, Texas, USA, January, 2015*. http://aaai.org/ocs/index.php/WS/AAAIW15/paper/view/10185

[49] Arnout Verheij, Allard Kleijn, Flavius Frasincar, and Frederik Hogenboom. 2012. A comparison study for novelty control mechanisms applied to web news stories. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on*, Vol. 1. IEEE, 431–436.

[50] Ellen M Voorhees. 2002. Overview of TREC 2002.. In *Trec*.

[51] Ellen M Voorhees. 2003. Overview of TREC 2003.. In *TREC*. 1–13.

[52] Charles L Wayne. 1997. Topic detection and tracking (TDT). In *Workshop held at the University of Maryland on*, Vol. 27. Citeseer, 28.

[53] Luca Weihs and Oren Etzioni. 2017. Learning to Predict Citation-Based Impact Measures. In *2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017*. 49–58. https://doi.org/10.1109/JCDL.2017.7991559

[54] Luca Weihs and Oren Etzioni. 2017. Learning to Predict Citation-Based Impact Measures. In *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*. IEEE, 1–10.

[55] Yiming Yang, Tom Pierce, and Jaime Carbonell. 1998. A study of retrospective and on-line event detection. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 28–36.

[56] Yiming Yang, Jian Zhang, Jaime Carbonell, and Chun Jin. 2002. Topic-conditioned novelty detection. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 688–693.

[57] Zaihan Yang and Brian D Davison. 2012. Venue recommendation: Submitting your paper with style. In *Machine learning and applications (ICMLA), 2012 11th international conference on*, Vol. 1. IEEE, 681–686.

[58] Yi Zhang, Jamie Callan, and Thomas Minka. 2002. Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 81–88.

[59] Pengfei Zhao and Dik Lun Lee. 2016. How Much Novelty is Relevant? It Depends on Your Curiosity. In *39th International ACM SIGIR Conference on Research and Development, Pisa, Italy*. 100.

[60] Xiaodan Zhu, Peter D. Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *JASIST* 66, 2 (2015), 408–427. https://doi.org/10.1002/asi.23179