Evaluating Research Novelty Detection: Counterfactual Approaches

Reinald Kim Amplayo and **Seung-won Hwang** and **Min Song**
Yonsei University
Seoul, South Korea
{rktamplayo, seungwonh, min.song}@yonsei.ac.kr

Abstract

In this paper, we explore strategies to evaluate models for the task research paper novelty detection: Given all papers released at a given date, which of the papers discuss new ideas and influence future research? We find the novelty is not a singular concept, and thus inherently lacks of ground truth annotations with cross-annotator agreement, which is a major obstacle in evaluating these models. Test-of-time award is closest to such annotation, which can only be made retrospectively and is extremely scarce. We thus propose to compare and evaluate models using counterfactual simulations. First, we ask models if they can differentiate papers at time t and counterfactual paper from future time $t + d$. Second, we ask models if they can predict test-of-time award at $t + d$. These are proxies that can be agreed by human annotators and easily augmented by correlated signals, using which evaluation can be done through four tasks: classification, ranking, correlation and feature selection. We show these proxy evaluation methods complement each other regarding error handling, coverage, interpretability, and scope, and thus altogether contribute to the observation of the relative strength of existing models.

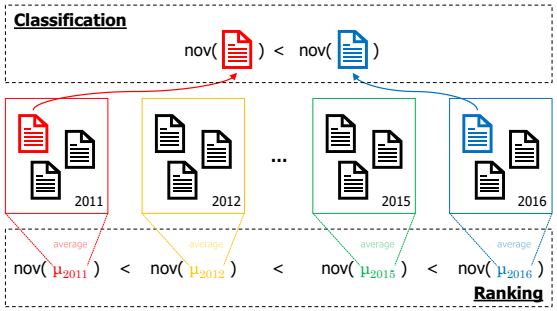
1 Introduction

Research paper novelty detection can be defined as follows: Given the full-text content of the paper, determine if the paper is novel or not. When comparing the novelty of two papers, we assume that only the texts (i.e., abstract, body, and reference sections) are shown and that both papers are published at the same time, and the venue is not known. This task is essential because while previous works on plagiarism detection (Harris, 2002; Lukashenko et al., 2007), citation recommendation (He et al., 2010; Ji et al., 2017), and reviewer assignment (Long et al., 2013; Liu et al., 2014) help in the administrative part of the review process, automatically detecting paper novelty can speed up the paper reviewing. Also, from the viewpoint of paper readers, it helps in filtering out non-novel papers from the large number of papers being published every day.

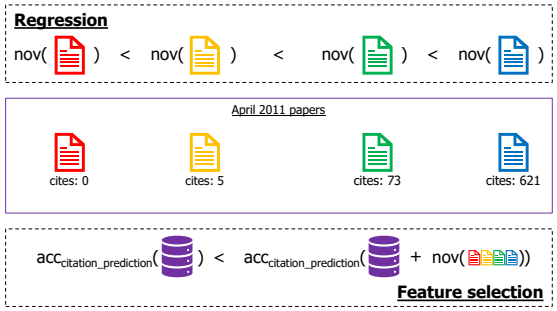
Despite its importance, this direction of research has not been explored as much. We argue that this is because it is hard to evaluate these models. An obvious solution is to create an evaluation dataset which contains papers that are labeled with their novelty. However, acquisition of this dataset is practically impossible, because of several aspects. First, novelty is not a singular concept, and captures diverse aspects, being *yet to be seen* with respect to previous knowledge and *innovative* to have *impact* on future publication. As a result, collecting novelty judgment from reviewers (e.g., from reviewing papers) as ground-truth, would have low cross-annotator agreement, especially when the qualification and background of reviewers are diverse. Second, one ideal solution is to create a dataset of test-of-time awarded papers, which are selected by widely-known highly qualified experts in the field, based on its impact after more than ten years since its publication. However, this process takes too long since we need to wait for ten years to determine if a paper stands the test of time or not.

In this paper, we explore evaluation methods which do not use gold labels for comparison, as similarly done in other NLP tasks: such as automated essay scoring (Burstein et al., 2004) and representation learning (Schnabel et al., 2015). Specifically, we consider the following counterfactual simulations, in place of human annotations: First, we ask models if they can differentiate papers at time t and counterfactual paper from future time $t + d$, where d is a large time gap. We found human annotators agree counterfactual pa-

Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13), pages 124–133
Hong Kong, November 4, 2019.



(a) Evaluation through time



(b) Evaluation through impact

Figure 1: Intuition behind our proposed evaluation metrics for research paper novelty detection, where $nov(\cdot)$ is the novelty detection model.

per from future is more novel, which suggest this *time proxy* simulation can capture one aspect of novelty. Second, we ask models if they can predict test-of-time award at $t + d$. As test-of-time award is a sparse signal, we augment with future citations, which we empirically observe to correlate with the award signal. Using this *impact proxy*, we evaluate whether the model can differentiate papers with high citations and those with fewer citations in the future. Through this, novelty detection can be treated as four NLP tasks: classification, ranking, correlation, and feature selection, as shown in Figure 1. Throughout the paper, we explain why and in what conditions we can evaluate models through these proxies.

2 Novelty Detection in Texts

Text novelty detection is a task to identify the set of relevant and noel texts from an ordered set of document swithin a certain topic (Voorhees, 2004). Text novelty is defined as providing new information that has not yet been found in any of the previously seen sentences. Most systems (Blott et al., 2004; Zhang et al., 2007) used TF-IDF metric as an importance value. Other novelty detection systems used entities such as named entities as features (Jaleel et al., 2004; Tsai and Zhang, 2011).

Unlike text novelty, research paper novelty is rather complex. While there is no clear and precise definition, Kaufer and Geisler (1989) attempted to describe research paper novelty into the points below:

1. Static: Novelty in a research paper is less a property of ideas than a **relationship** among research communities and ideas. It is less an individual trait than the consistency of the paper to the research community and structure.

2. Dynamic: Novelty in a research paper is defined by how much the introduction of this paper "changes" the overall relationship.

   - Research papers are novel if they **mastered** a set of conventions which enables the research community to use past ideas and go beyond in the future.
   - Novelty in a research paper is a shorthand for the standards used to contribute to the **growth** of a specific disciplinary research community.
   - Novelty in a research paper is a **balance** between the inertia of past ideas and the drive to contribute to the ideas.

The above description tells us that novelty in normal texts and research papers are not the same. While in normal texts, it is necessary that there is new information to be considered novel, in research papers, the relationship among various research entities (e.g., authors, ideas, contents, etc.) are more important, as also shown in the literature (Ding et al., 2013, 2014; Song and Ding, 2014; Amplayo and Song, 2016). We thus argue, and show in our experiments, that novelty detection models that use entity-based features are more effective for research paper novelty detection.

3 Novelty Detection Models

Most models for novelty detection can be described in two parts, First, a feature extraction module is used to select useful features. Second, a novelty scoring module is used to output a score that represents how novel the text is.

3.1 Feature Extraction Modules

We summarize the feature extraction methods in Figure 2. There are two types of features: Normal text features are features that are commonly

used in novelty detection for text which are not research papers. Citation features are features that make use of citation information (i.e., the reference section) that are usually available in research papers.

Normal Text Features Feature extraction from normal text include TF-IDF features (Blott et al., 2004) and word co-occurrence features (Gamon, 2006). Since these methods are primarily used to detect novelty in normal texts, they do not consider the existence of the relationship between citing and cited papers.

- `tfidf` (Blott et al., 2004; Zhang et al., 2007): product of the term frequency and the logarithm of the inverse of the relative document frequency.

- `cooccur` (Gamon, 2006): transforms text into a graph using the fact that two words within a window size are connected with an edge and extracts twenty-one features based on the node and edge statistics (e.g., number of new edges, ratio of node to edge, etc.).

Citation Features Feature extraction from citation information uses the idea that there is a direct relationship between the cited paper and the citing paper (Amplayo et al., 2018). Instead of using co-occurrence graphs, as in `cooccur`, these models use citation graphs to extract the features from the paper, where information from cited and citing papers are connected using a directed edge. There are two kinds of citation features: (a) **Metadata-based** citation features, in which we create an edge between two metadata information such as authors and papers, and (b) **Entity-based** citation features (Amplayo and Song, 2016), in which we create an edge between two entities extracted from the text content, such as keywords, latent topics, and words. Features are then obtained using the same method as in (Gamon, 2006).

- `paper`: simple citation graph where the nodes are papers and the edges are citation relationships between the papers: If paper a cites paper b then an edge from b to a exists.

- `author`: simple citation graph where the nodes are authors and the edges are citation relationships from the authors of the cited paper to the authors of the citing paper.

- `keyword`: a citation graph where the nodes are RAKE-generated keywords (Rose et al., 2010) extracted from the citing and cited information[1], and the edges are connected from the cited keywords to the citing keywords.

- `topic`: a citation graph where the nodes are LDA topic vectors (Blei et al., 2003) extracted from both the citing and cited information, and the edges are connections from the cited to citing topics.

- `word`: a citation graph where the nodes are lowercased and lemmatized nouns from both citing information and cited information.

3.2 Novelty Scoring Module

The features extracted by the models described above can be compared over a common novelty detector. Majority of previous work use autoencoders, neural networks that are used to learn efficient codings in an unsupervised manner. When used as a novelty detector (Japkowicz et al., 1995), autoencoders encode the input features x into encodings and try to decode an output x' such that x and x' are equal.

This idea can also be leveraged to detect research paper novelty as well. First, we train the autoencoder using features extracted from papers of the training data. The papers in the training data are papers from the past and represent the current known research ideas and communities. Then at test time, for each unseen paper p, we extract its input feature x_p using the feature extraction module. The autoencoder accepts x_p as input, and outputs x'_p. The novelty score of the paper is the closeness of x_p and x'_p, which is calculated as the root of the sum of the squared difference of x_p and x'_p. If x_p and x'_p are nearly identical, then the features extracted from paper p have already been expected by the model; thus p is not novel. Otherwise, p contains new information and hence is considered novel. The autoencoder is then retrained to include the current paper p for the next unseen paper.

4 Dataset of Research Papers

We gathered computer science research papers as an evaluation dataset from arXiv, an online repos-

[1]Hereon, the citing information refers to the abstract of the paper, while the cited information refers to the snippet containing the corresponding in-text citation

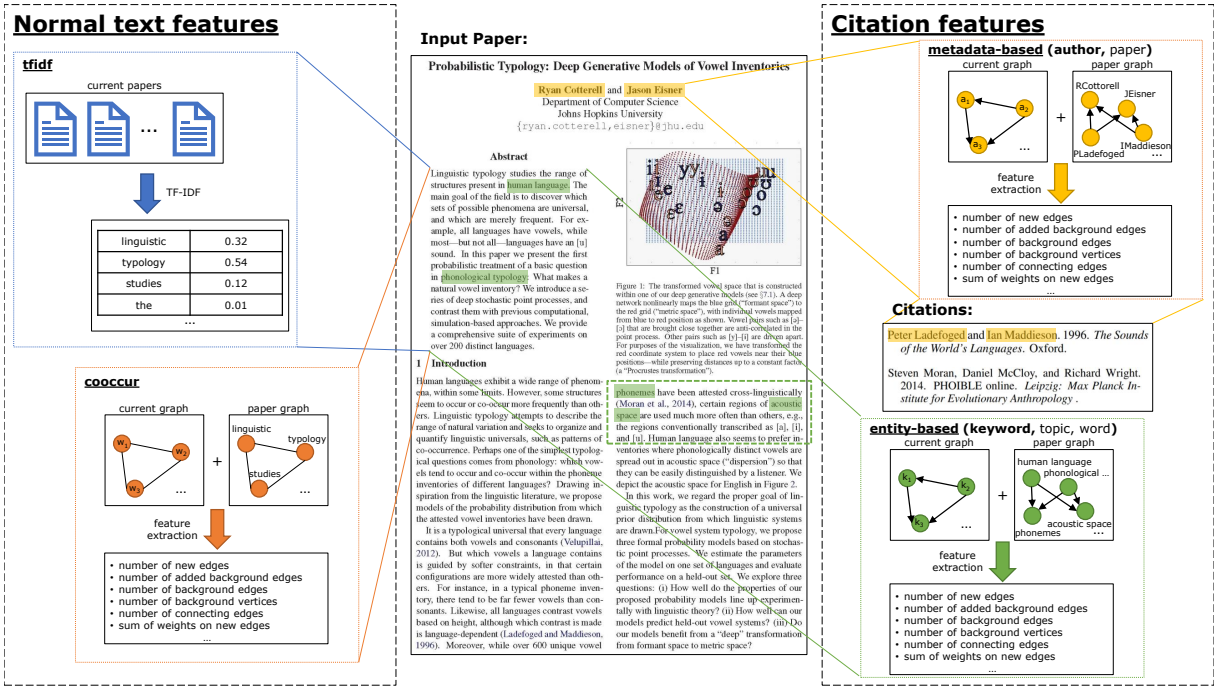


Figure 2: The feature extraction methods used by different novelty detection models. We only show `author` and `keyword` as examples for metadata-based and entity-based citation features, respectively, for conciseness.

itory that covers a wide range of computer science subfields and has a total of forty subcategories, ranging from Artificial Intelligence to Systems and Control. This repository is a good source of a variety of papers that are accepted or rejected by conferences and journals, and may contain ideas already presented in the past. Each paper has its information such as the title, author names, submission date and abstract, and its full text in PDF format. We additionally collected the published date information and citation count information from the same tool. The data gathered consists of papers from the year 2000 to 2016. We divided the data into seed data (year 2000-2005)[2], training data (year 2006-2010) and test data (year 2011-2016).

5 Counterfactual Simulation on Time

We first evaluate research paper novelty detection models using counterfactual simulations on time. The intuition behind this idea is that, when papers from t and $t+d$ are presented to human annotators, they would agree on the latter to be more novel (Section 5.1). Once this holds, we can evaluate as classification between t and $t+d$ (Section 5.2), or ranking (Section 5.3) among clusters of papers ground by a specific time period.

5.1 Preliminary analysis

To test our intuition, we performed a preliminary experiment as follows. We collected pairs of similar papers from ACL 2011 and ACL 2016; we used similarity scores from averaged pretrained word2vec word vectors (Mikolov et al., 2013) to pair the papers. We asked four senior graduate students who major in natural language processing to judge which among the pair is more novel or not. To perform a transparent experiment, we did not reveal the data collection process and only show the title and the abstract of the papers. We gave annotators three choices: (A) choose paper A, (B) choose paper B, and (C) no idea.

The results show that 42.5% of the time the graduate students selected the 2016 papers, while the 2011 papers were selected only 8.5% of the time. This indicates that time can be used as a proxy for novelty detection. In the remaining 49.0% of the time however, the students had no idea which paper is more novel. We posit that this is because they are only experts on a small sub-area of NLP, which additionally shows the difficulty of creating an expert-annotated evaluation set for the problem.

[2]The seed data is needed by models using citation features (Amplayo et al., 2018). Other models use both seed and training data for training.

Model	Accuracy
tfidf	55.56
cooccur	56.49
author	59.46
paper	59.48
word	59.90
keyword	61.11
topic	**78.19**

Table 1: Accuracies of different models on the time classification task.

Model	1 year	6 months	3 months	1 month
tfidf	-0.029	-0.100	-0.099	-0.001
cooccur	0.771	0.682	0.549	0.384
author	0.543	0.536	0.362	0.177
paper	0.771	0.882	0.859	0.821
word	**1.000**	0.982	0.911	0.835
keyword	**1.000**	0.945	0.930	0.863
topic	**1.000**	**0.991**	**0.986**	**0.979**

Table 2: Ranking results. Spearman correlation coefficient ρ for each period each novelty detection models in comparison. Bold-faced numbers are the best scores.

5.2 As a Classification Task

Intuition As mentioned above, the first condition is that the difference between publication times of non-novel and novel papers should be large. Using this, we can reduce the research paper novelty detection task into a classification task, where the task is to classify if a paper is novel or not. Given a large time difference d, the papers published in the past time t are given a "not novel" label while papers published in time $t+d$ are given a "novel" label. If a classification model trained using the novelty scores extracted from a novelty detection model can easily distinguish papers from different groups, then the model performs well.

Setup We use the 2011 test data as papers published in time t and the 2016 test data as papers published in time $t + d$, where the time difference d is five years. Moreover, we extract features using novelty detection models trained only until the year 2010 training data. That is, we do not retrain using 2011 papers to extract features from 2016 papers. We then train a logistic classifier using the extracted features as input and the novelty label as output. We evaluate the classifier by calculating its accuracy using 10-fold cross validation. We report the average accuracy of the ten subsamples. Note that since the classifier does not have information regarding the publication date, it tries to classify papers knowing that all of them are published at the same time (i.e., end of 2010).

Results The classification accuracies are reported in Table 1. Results show that among the competing models, `topic` performs the best in terms of accuracy. All entity-based models perform better than other competing models, and among these models, `word` performs the worst. This is because some words may not carry novel information specific to the research field. The metadata-based models, `author` and `paper`, perform similarly but perform worse than the entity-based models. Finally, `tfidf` and `cooccur` perform the worst among all models. Note that the lower bound accuracy where we assume all papers are not novel is 55.56%. `tfidf` model performs eqaully with the lower bound, which shows that the classifier trained using TF-IDF features is not able to distinguish novel and non-novel papers.

5.3 As a Ranking Task

Intuition The second condition mentioned above is that evaluation through time should be done in clusters of papers, instead of individual papers, which are grouped by publication time. Using this condition, we can reformulate the novelty detection task into a ranking task, where we are tasked to rank the clusters by their publication time, using the average novelty score of papers in the cluster. Formally, let $P_1, P_2, ..., P_n$ be the sequential sets of papers grouped by given multiple publication time period $t_1, t_2, ..., t_n$. Then, the average novelty scores of the papers in the groups should be sequentially increasing, i.e., $\mu(P_1) < \mu(P_2) < ... < \mu(P_n)$, where $\mu(P)$ stands for the average novelty scores of research papers in set P. We can then use a correlation function to measure the monotonicity between the relationship of the period and the mean of the papers in that period.

Setup In this setup, we use four types of publication time periods, i.e., 1 year, 6 months, 3 months, and 1 month. We extract features using models trained until the year 2010 training data to view all papers as papers published at the same time. Through this setup, the model does not have information regarding the publication date. Hence it ranks papers knowing that all of them are published at the same time (i.e., end of 2010). We use all the test data from years 2011 to 2016 and divide them according to the different periods. We

Scenario	BEST #cites	ave. #cites	% below
A	210	76.8	92.0%
B	34	25.0	86.3%

Table 3: Results of preliminary analysis for the impact proxy. The BEST refers to the best paper.

calculate the Spearman rank-order correlation coefficient ρ as the rank-order correlation function.

Results The results are shown in Table 2. Overall, the entity-based models perform the best among all competing models. Interestingly, all entity-based models show a perfect correlation coefficient when using a 1-year time period. Among the entity-based models, `topic` performs the best on all periods. One interesting result is that as the period gets smaller (i.e., from annually to monthly), the correlation coefficient gets smaller. We argue that this is because as the period gets smaller, the evaluation condition above gets weakened and thus evaluation gets unreliable. The `tfidf` produces a negative correlation, which means the novelty scores are decreasing and thus contradicts to the intuition described above.

6 Counterfactual Simulation on Impact

Another way to evaluate research paper novelty detection models is performing counterfactual test-of-time prediction at time t. However, award annotation at $t+d$ is sparsely annotated to very few papers. We observe the number of citation counts the paper is correlated with award prediction, to augment citation for impact proxy. As a specific example, if both papers a and b are published at the same time, and paper a received more citations than paper b, then paper a is more novel than b. Through this, we can evaluate models both intrinsically by assuming the number of citation count correlates with the novelty of the paper (Section 6.2) and extrinsically by using the novelty scores as features for the citation count prediction task (Section 6.3).

6.1 Preliminary Analysis

To test our idea, we performed a preliminary experiment to check if novelty correlates with citations with the condition that they are published at the same time. We considered papers that received best paper awards as a rough estimate of a paper that is more novel than papers published at the same time. We looked at two different scenarios: (A) ACL 2011 best paper versus other ACL 2011 papers (total of 163 papers), and (B) ICML 2011 best paper versus 300 arXiv papers[3] with upload dates closest to the ICML best paper.

We compared the best paper and other papers using two methods. First, we calculated the average number of citations of other papers and compared it with the number of citations of the best papers. Second, we also calculated the percentage of papers with citations below the number of citations of the best paper. Results are shown in Table 3. Results show that in both scenarios, both papers have higher citations than average. Moreover, at least 86.3% of the papers have a lower number of citations than the best paper. This shows that citations can be used as a proxy for novelty detection.

Model	1-Month	3-Month	5-Month	AVG
`tfidf`	-0.043	-0.013	0.014	-0.014
`cooccur`	0.066	0.098	0.018	0.060
`author`	0.070	0.128	0.023	0.074
`paper`	0.079	0.097	0.034	0.070
`word`	0.123	0.076	0.045	0.081
`keyword`	**0.332**	**0.461**	**0.271**	**0.355**
`topic`	0.137	0.204	0.093	0.145

Table 4: Correlation results. Pearson correlation coefficient between different novelty detection models and citation counts. Bold-faced values are top three values.

6.2 As a Correlation Task

Intuition We use citation counts as labels to perform evaluation, assuming that the novelty score correlates with the citation count of the paper. One condition must be considered: Papers should already be *mature*, that is, enough time should have been given to the paper to be exposed to the research community. Using these assumptions, we can simplify the task into a correlation task where we check the relationship between the novelty scores and the citation counts.

Setup To assure paper maturity, papers published recently are not used for evaluation. Instead, we use papers that are published approximately five years before the time data gathering was done, i.e., since we gathered the data on April 2016, we use papers that are published near the April 2011 date. We consider three sizes of windows: 1-month window where we consider only April 2011 papers, and 3/5-month windows where we can consider March to May 2011 papers and

[3]We use arXiv papers for comparison not only for the diversity of scenarios but also since most ICML 2011 papers (including the best paper) were uploaded to arXiv before the conference started.

February to June 2011 papers, respectively. We use the Pearson correlation coefficient as the evaluation metric. We also report the average correlation.

Results We show the correlation coefficient scores of all competing models in Table 4. The table shows that the entity-based models outperform all the other models. Among them, `keyword` performs the best on all window sizes. `tfidf` performs the worst among the models, garnering negative correlation when the window size is 1/3-month, which means that the novelty scores produced by the model do not correspond to the citation count.

6.3 As a Feature Selection Task

Intuition We evaluate the novelty detection models by measuring their contribution to the citation count prediction task, where we are given a paper and its information, and we are tasked to predict the number of citations the paper receive after a particular given time. Previous works have attempted to use content features (Yan et al., 2011; Chakraborty et al., 2014) to solve the task. We argue that the novelty scores can also be treated as content features, and the novelty detection model that produces the most useful content feature can be regarded as the best model for the citation count prediction task. Using this intuition, we can reformulate the task as a feature selection task.

Setup Following Yan et al. (2011), we perform a feature selection study on the novelty scores produced by all competing models. Specifically, we look at the performance of a citation count prediction model, both (a) when only one novelty score is isolated as a single feature, and (b) when the same novelty score is dropped and all other novelty scores are used as features. We use five typical models for citation count prediction: linear regression (LR), k-nearest neighbors (KNN), decision trees (DT), support vector machines (SVM), and multilayer perceptron (MLP). To train these models, for each year from 2011 to 2015, we use the first five months as training data, and the sixth month as test data, obtaining five training and test datasets. We use the coefficient of determination R^2 as the evaluation metric. Finally, we also report the average of results on the five datasets.

Results The results are shown in Table 5. `keyword` performs the best among all models, having scores included in the best three scores for all cases. `topic` also performs well, having nine out of ten scores in the top three. Interestingly, both `author` and `word` perform comparably, having six and five out of ten scores in the top three, respectively. We posit that this is due to previous findings (Yan et al., 2011; Chakraborty et al., 2014) that author-based features (e.g., h-index, author rank) are informative when predicting citation counts. `author` may have learned author-specific biases in its novelty scores. Finally, `tfidf` performs the worst, consistently having a zero score on all classifiers when used in isolation. This means that `tfidf` does not carry any important signal to predict citation counts.

7 Discussions

The evaluation metrics are complementary to each other We compared existing models using four evaluation methods. These evaluation methods have different characteristics that complement each other. For one thing, evaluation through time prefers topically new papers, thus evaluation would be problematic on published papers that are off-topic or published at a low-tier conference or journal. However, this problem would not exist when evaluating through impact, because off-topic and low-tier papers normally do not have lots of citations. Moreover, evaluation through citation prefers papers that have many citations, hence it can be problematic on less impactful yet highly cited papers, such as survey papers. This is not a problem when evaluating through time, because the topics discussed in survey papers are not new.

The individual tasks also complement each other. For example, the classification task provides a direct comparison of individual paper novelty since evaluation is done at paper-level, while the ranking task is not very interpretable since evaluation is done in groups. However, the classification task is not able to evaluate papers between two periods t and $t + d$, while the ranking task makes use of all data. It is therefore recommended that all four evaluation methods are used for evaluation.

Novelty in normal texts and research papers are different The results from different evaluation methods presented in the paper consistently show that models which are originally used for detecting novelty in normal text, especially `tfidf`, do not perform well in the task. This contradicts to previously reported strong results (Voorhees,

Classifier	tfidf	cooccur	author	paper	word	keyword	topic
when feature is isolated							
LR	0.000	0.007	**0.091**	0.082	0.042	**0.104**	**0.086**
KNN	0.000	0.007	**0.010**	0.002	0.003	**0.027**	**0.015**
DT	0.000	0.002	0.000	0.000	**0.038**	**0.023**	**0.003**
SVM	0.000	0.007	**0.084**	0.082	0.042	**0.104**	**0.086**
MLP	0.000	0.006	**0.100**	0.083	0.075	**0.091**	**0.090**
when feature is dropped							
LR	0.110	0.110	0.111	0.120	**0.109**	**0.083**	**0.108**
KNN	0.038	0.044	0.045	0.038	**0.036**	**0.035**	**0.026**
DT	0.067	0.067	**0.039**	0.067	**0.000**	**0.055**	0.060
SVM	0.114	0.115	**0.107**	0.119	0.113	**0.098**	**0.104**
MLP	0.107	0.105	0.107	0.105	**0.103**	**0.104**	**0.098**

Table 5: Feature selection results. R^2 score of each feature (novelty score) when in isolation and when dropped from an all-features model. When in isolation, the larger R^2 is, the better. When dropped, the smaller R^2 is, the better. Bold-faced values are top three values per row.

2004) on normal text novelty detection task, where the model obtained state of the art performance. This supports the fact that the novelty in normal texts and in research papers is different. Since information in a research paper is often not completely new, the `tfidf` model always gives a small novelty score to newer papers, which explains the negative results.

Entity-based citation graphs are better feature extractors On all evaluation metrics, models with features extracted from entity-based citation graphs, i.e. `keyword` and `topic` obtain the best performance. One possible explanation to this is that features extracted from entity-based citation graphs best reflect the changes brought upon by the paper on the relationship between entities in the background knowledge (Ding et al., 2013). This is defined as the static and dynamic nature of research paper novelty in Section 2. On the other hand, normal text novelty detection models capture features that are different from research paper novelty, and metadata-based citation graph models capture only the static nature of research paper novelty.

Limitations in models and evaluation One major limitation of the existing methods is that they do not consider the possible existence of biases from other factors. For example, an application of a basic machine learning model on other non-computing fields such as philosophy or psychology may have different intensities of novelty depending on the field of the venue where the paper is published. Also, papers from multidisciplinary fields can receive more citations because two or more research communities are reading them. These factors may affect the evaluation presented in this paper.

Regarding evaluation, expert-annotated labels, though difficult to acquire in scale, are still more desirable than our suggested estimates. The ideal way to create this dataset is to gather papers evaluated for the *test of time award*, an award given to papers published in the past 10 or more years for their significant contribution by a committee of influential researchers of the community. This ensures that the annotators are experts and the papers have shown a noticeable impact. However, this method is not efficient as it will need at least ten years to annotate a small number of papers. Moreover, this does not include annotations of papers with less intensity of novelty.

8 Conclusion

We described counterfactual simulations to evaluate the research paper novelty detection task by using time and impact as proxies. Using these methods, we evaluate features used in existing models, to find entity-based citation features, compared to normal text features, are stronger signals to predict and explain novelty. We finally provided discussions on the advantages and disadvantages of using these evaluation metrics and the future direction of this research.

Acknowledgement

This work is supported by Micrsoft Research Asia.

References

Reinald Kim Amplayo, SuLyn Hong, and Min Song. 2018. Network-based approach to detect novelty of scholarly literature. *Inf. Sci.*, 422:542–557.

Reinald Kim Amplayo and Min Song. 2016. Building content-driven entity networks for scarce scientific literature using content information. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining, BioTxtM@COLING 2016, Osaka, Japan, December 12, 2016*, pages 20–29.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Stephen Blott, Fabrice Camous, Paul Ferguson, Georgina Gaughan, Cathal Gurrin, Gareth J. F. Jones, Noel Murphy, Noel E. O'Connor, Alan F. Smeaton, Peter Wilkins, Oisín Boydell, and Barry Smyth. 2004. Experiments in terabyte searching, genomic retrieval and novelty detection for TREC 2004. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–36.

Tanmoy Chakraborty, Suhansanu Kumar, Pawan Goyal, Niloy Ganguly, and Animesh Mukherjee. 2014. Towards a stratified learning approach to predict future citation counts. In *IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014*, pages 351–360.

Ying Ding, Min Song, Jia Han, Qi Yu, Erjia Yan, Lili Lin, and Tamy Chambers. 2013. Entitymetrics: Measuring the impact of entities. *CoRR*, abs/1309.2486.

Ying Ding, Guo Zhang, Tamy Chambers, Min Song, Xiaolong Wang, and Chengxiang Zhai. 2014. Content-based citation analysis: The next generation of citation analysis. *JASIST*, 65(9):1820–1833.

Michael Gamon. 2006. Graph-based text representation for novelty detection. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 17–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Robert Harris. 2002. Anti-plagiarism strategies for research papers. *Virtual salt*, 7.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and C. Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 421–430.

Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. Umass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Nathalie Japkowicz, Catherine Myers, and Mark A. Gluck. 1995. A novelty detection approach to classification. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal Québec, Canada, August 20-25 1995, 2 Volumes*, pages 518–523.

Xiaonan Ji, Alan Ritter, and Po-Yin Yen. 2017. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *Journal of Biomedical Informatics*, 69:33–42.

David S Kaufer and Cheryl Geisler. 1989. Novelty in academic writing. *Written Communication*, 6(3):286–311.

Xiang Liu, Torsten Suel, and Nasir D. Memon. 2014. A robust model for paper reviewer assignment. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014*, pages 25–32.

Cheng Long, Raymond Chi-Wing Wong, Yu Peng, and Liangliang Ye. 2013. On good and fair paper-reviewer assignment. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 1145–1150.

Romans Lukashenko, Vita Graudina, and Janis Grundspenkis. 2007. Computer-based plagiarism detection methods and tools: an overview. In *Proceedings of the 2007 International Conference on Computer Systems and Technologies, CompSysTech 2007, Rousse, Bulgaria, June 14-15, 2007*, page 40.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. pages 1–20.

Tobias Schnabel, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 298–307.

Min Song and Ying Ding. 2014. *Topic Modeling: Measuring Scholarly Impact Using a Topical Lens*, pages 235–257. Springer International Publishing, Cham.

Flora S. Tsai and Yi Zhang. 2011. D2S: document-to-sentence framework for novelty detection. *Knowl. Inf. Syst.*, 29(2):419–433.

Ellen M. Voorhees. 2004. Overview of TREC 2004. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*.

Rui Yan, Jie Tang, Xiaobing Liu, Dongdong Shan, and Xiaoming Li. 2011. Citation count prediction: learning to estimate future citations for literature. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 1247–1252.

Kuo Zhang, Juan Zi, and Li Gang Wu. 2007. New event detection based on indexing-tree and named entity. In *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*, pages 215–222.