Impact factor analysis report of paper acceptance at ICLR

Shuangyang Hu

Master of Commerce, University of Sydney, NSW 2008, Australia

Email: shhu2848 @uni.sydney.edu.au.com

Abstract. In this paper, Visual Data Analytics (VDA) approaches have been used to analyze the factors influencing the acceptance of a scientific paper to a top representation learning conference. This shall help authors determine the likelihood of their paper being accepted by the leading representation learning conferences. This report uses the International Conference on Learning Representations (ICLR) dataset, which contains publicly available documents collected at ICLR conferences from 2017 to 2021. This report uses visualization to analyze the impact factors of acceptance to find out. The novelty of the paper is that it analyzes the impact factor of whether a scientific paper will be accepted at a top representation learning conference, laying the groundwork for future predictions of scientific paper acceptance.

Keywords: Visual Data Analytics, Deep Learning, Data Science, Artificial intelligence, Machine Learning, Paper Review, Classification, Acceptance prediction

1 Introduction

1.1 Background

The International Conference on Learning Representations (ICLR) is a machine learning conference that includes invited talks as well as oral and poster presentations of refereed papers. ICLR focuses on presenting and publishing cutting-edge research on deep learning used in the fields of artificial intelligence, statistics and data science, but also in important application areas such as machine vision, computational biology, speech recognition, text item description, etc. [2] Along with ICML and NeurIPS, ICLR is one of the three major machine learning and artificial intelligence conference, and has the highest impact of the three. [1]

1.2 The significance of the topic

There is now a significant increase in the number of conferences held worldwide, and a substantial portion of these conferences are related to machine learning and artificial intelligence (AI), which is currently a relatively active area of research in computer

V. Escudero et al. (Eds.): BADS 2022, AHCS 11, pp. 657–669, 2023.
https://doi.org/10.2991/978-94-6463-102-9_68

science. With the popularity of machine learning applications, a wide variety of related scientific papers have also emerged. This will be a rare opportunity for researchers in related fields.

1.3 Research aims and objectives

The goal is to facilitate scholars in the field of machine learning and artificial intelligence to understand the central topic requirements and the quality of manuscripts needed to publish at prestigious conferences.

To achieve this goal, this report uses Visual Data Analytics (VDA) to analyze the impact factor of a paper's acceptance at a top machine learning conference. Although machine learning techniques have been widely used in various fields, there is still limited research published on the subject, which has somewhat limited scholars' ability to translate their research results into academic papers.

1.4 Overview of report structure

This report is organized as follows: in section II, a brief overview of the literature review is presented. in section III, the methodology of the study is described. In section IV, the results of the study are presented. Finally, in Section V, the paper concludes with a discussion of future work.

2 Literature Review

2.1 Overview of the literature

The article by Kang, D. presents the first scientific peer-reviewed public dataset (PeerRead v1) that can be used for research purposes and based on this analysis predicts the acceptance of papers based on textual features and predicts scores based on each aspect of papers and reviews. A high correlation between overall recommendation and recommendation of oral presentations was found, and having an appendix would have a higher acceptance rate.[5]

The article from Ghosh, A. focuses on research testing the hypothesis that the non-core content factors of research papers can indicate their acceptance potential. The data from ICLR 2017 papers were studied using 14 supervised learning models and 5 unsupervised. The words "performance", "architecture", "method", and "good" were the top-ranking words, which indicates that reviewers are very concerned about these things and expect publications with "excellent performance", good/better "architecture" and novel "methods", and finally the models they have constructed with an accuracy rate of 65%. [3]

The paper from Momen, S. uses ML models to predict the acceptance of papers in top AI recalls as a reference for computer science and artificial intelligence researchers. The highest ACCEPTANCE was eventually obtained for RANDOM FOREST, with an accuracy of 81%.[6]

In this paper from Wang, 2021, in order to predict paper acceptance at the institutional level, this work formalizes the problem as a regression task a set of factors that determine paper acceptance is explored. Affiliation scores are first calculated to construct a collaborative network of institutions and to analyze the importance of institutions using a network centrality metric. Then four measures of authors' influence and competence are extracted to consider authors' contributions. Finally, a random forest algorithm is used to solve the prediction problem of paper acceptance. As a result, this paper improves the ranking of paper acceptance rate NDCG@20 to 0.865, which is better than other state-of-the-art methods.[8]

The paper from J. Joshi predicts whether a paper will be accepted for a conference by combining machine learning clustering algorithms and natural language processing. The model makes predictions based on the extracted features. The features considered for most conferences are the number of references, bag of words for ML related terms, etc. The model was trained on the above dataset, containing 70 accepted and 100 rejected papers. The model is trained by applying algorithms such as logistic regression, decision tree, and random forest for prediction. A comparative study shows that Decision Tress works effectively by providing 85% accuracy.[4]

2.2 Limitations analysis

The limitations of the above five articles are: 1) Focus on the study of model structure and model testing, not enough description of the feature. 2) Lack of borrowing more advanced NLP and machine learning models to analyze plagiarism and articles and other data on a larger scale. 3) Limitations of the data material.

Based on this, this paper thinks a clear report showing features is needed before post predictive modelling. Unlike these five papers, this paper will focus on using Visual Data Analytics (VDA) to analyze the impact factor of a paper's acceptance at a top machine learning conference. This will build a good foundation for subsequent modelling studies.

3 Methodology

3.1 Scope of the study

For the impact factor analysis of ICLR conference accepted papers, this paper uses the affiliation and paperlist data tables of ICLR for 2017-2021, with a total of 10 tsv. These two datasets are presented separately below.

3.2 Data processing

1) Paperlist data related.

a) Processing data part.

After reading the data and checking the basic information, this paper found that all the data have null values. There are redundant symbols in many places of the data, and each data set has different kinds of describe. So, the processing direction is initially divided into processing null values, removing redundant symbols and integrating the types of decision data.

Since 'one-sentence_summary' has a large number of missing values, this paper considers this part of the data not to have a high reference value and is dropped since the missing values in the decision part of the data for 2017 and 2020. The decision is the statistics of accept data, so this part of the missing value is supplemented with 'Not Given'.

Take the 2017 data as an example, 'keywords', 'authorids', 'Rating', 'authors' have '[' and ']' which are not needed for data analysis, so they were removed. (The same treatment was done for the remaining years.)

Since this paper is studying the impact factor of paper acceptance, the processing of 'decision' data is particularly important. To study the acceptance situation, so the initial idea is to divide the data into accept and reject, so that the acceptance rate can be well obtained for subsequent analysis. In this regard, each dataset's specific types of the decision were counted. Taking 2017 as an example, the data were divided into 'Reject', 'Accept (Poster)', 'Not Given', and 'Accept (Oral)', 'Invite to Workshop Track' and 'Accept' were categorized as 'Accept' and 'Reject' were grouped into two sections. (The same treatment was done for the remaining years.)

b) Integrating data part.

As mentioned earlier, some data come with meaningless and unwanted symbols, which are removed and organized in this pair.

The 'Keywords' part uses the data of 2021 and finds that there is a line consisting of multiple keywords and interval symbols, to facilitate the subsequent analysis, this paper composes all the strings in a Series into a long string, then into a list. And there is still, ' symbols and spaces in the string, so use the replace function to remove them. After taking out and processing to get a long list, it is found that there is the same word case inconsistency, which will affect the categorization of keywords, the same word may be divided into different categories because of the case inconsistency. So, use the title function to set all the first letters to uppercase and the rest of the letters to lowercase by default. This is integrated to get a complete frequency statistics table of 'Keywords' in 2021.

The 'Rating' section also uses the data from 2021, and inspection reveals the 'Rating' data, which is composed of multiple numbers, symbols, and words. It will be processed into the average rate for subsequent analysis and research and added to the table in this paper. First, as a comma is also part of rating comments, this report uses a particular char # for splitting. Subsequently, each second character x[number] is a score, so int() is added to convert char 'number'. Finally, these values are added to the list of scores. Since an exact number is needed for the analysis, it is added to the paperlist of 2021 using the mean value and naming it 'average_scores'.

2) Affiliation data related.

After reading the data from 2017-2021 and performing a merge of the 5-year data tables, then showing the underlying information of the data, it was found that there was a null value in the 'decision' data. So, the fillna function was used to fill in the null value as 'Not Given'.

4 Results& Discussion

The center of this paper is the impact factor analysis of ICLR conference accepted papers, where the acceptance rate of a paper is influenced by the topic, publication channel, and reviewers' comments. [7] Based on the cleaned and integrated data and the second part of the Literature Review, this paper will study the relationship between accept rate and keywords, rating, and organization.

4.1 Paperlist data related

1) About accept rate.

a) Accept and Reject.

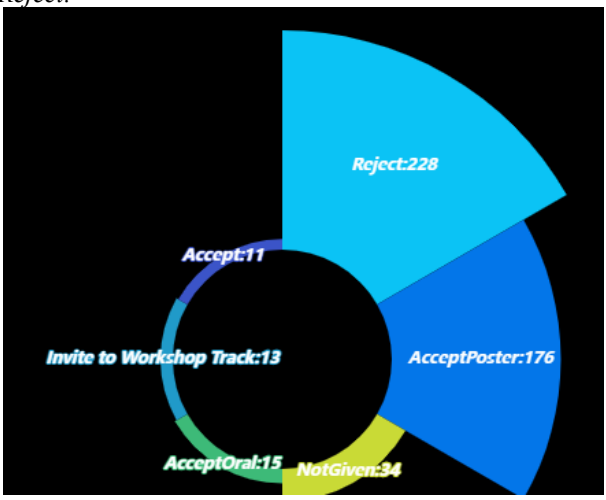

Fig. 1. Nightingale rose diagram of the decision data distribution in 2017

Firstly, for the 2017 decision data, statistics were conducted, the maximum number of 'Reject' was 228, 'Not Given' was 34, and 'Accept' was the least at 11. However, since these parts can be combined, they are unified as 'Accept' and 'Reject' to facilitate the subsequent calculation of accept rate.

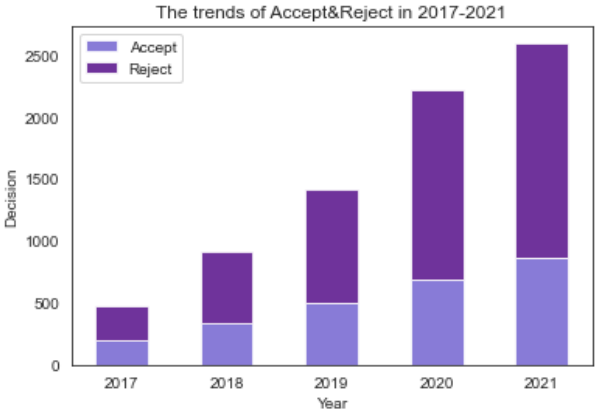

Fig. 2. Histogram of the accept and reject trends in 2017-2021

Second, the five-year data from 2017-2020 after unified classification is displayed, the decision can represent the number of manuscripts accepted by ICLR. As a whole, this graph illustrates that the number of submissions has been increasing during the five years, and along with the growth of the number of submissions, the number of accepts and rejects also shows a growing trend. This paper argues that this indicates that the influence of ICLR conferences is increasing on the one hand, and that more and more machine learning scholars are writing papers for publication on the other. The overall number of rejects, which represents the dark color, is higher than accept, which indicates that although the number of submissions is increasing, the quality is still lacking. It also shows that it is relevant to study the impact factor of accepted papers in ICLR conferences.

b) Accept rate.

In order to further study the change of accept between five years, this paper calculated the acceptance rate of each year according to the formula

$$Accept\ Rate = Accept/(Accept + Reject) \tag{1}$$

and drew a line graph.

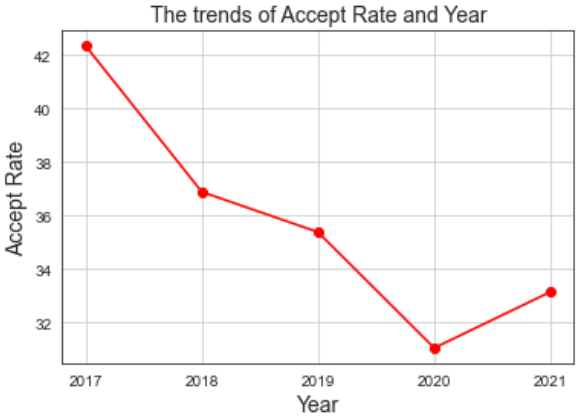

Fig. 3. Line chart of Accept rate trends in 2017-2021

The graph shows an overall decreasing trend, dropping to the lowest in 2020, but rebounding in 2021. This further visualizes the conclusion obtained above: the number of submissions is rising, but the quality is still lacking, and it is highly informative for machine scholars for us to do such a study.

2) About keywords.

Based on the data processing in the previous section, the data of 2021 was selected for the study. First, the first 160 keywords in 2021 were plotted as WordCloud.

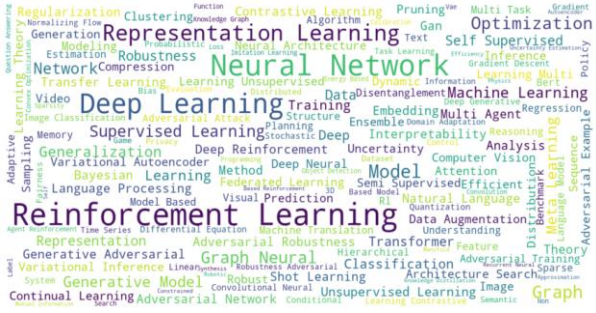

Fig. 4. World Cloud of the most frequently occurring keywords in 2021

WordCloud shows the keywords as words of different sizes according to their frequency of occurrence. The figure shows that 'Deep Learning' and 'Reinforcement Learning' appear more frequently.

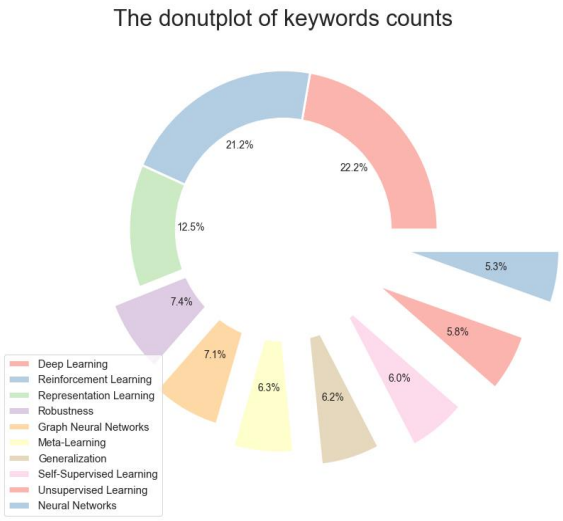

Fig. 5. Donut plot of the percentage of top 10 keywords appearing in 2021

However, since WordCloud cannot represent the proportion of keywords with high frequency, the Donut-plot is plotted in this paper to investigate the ten keywords with the highest frequency. Among these ten words, 'Deep Learning' (22.2%) and 'Reinforcement Learning' (21.2%). This data can indicate the general topics submitted by scholars, since it is not combined with the acceptance rate. It does not indicate that represents a trend in ML or that papers will be accepted by ICLR if they are related to these words.

3) About keywords and accept rate.

The data in this section are selected from the top seven keywords in 2021. The first keyword was 'Deep learning', and the acceptance rate was calculated (the same process was done for other keywords). The relationship between the top seven keywords and the acceptance rate in 2021 is plotted.

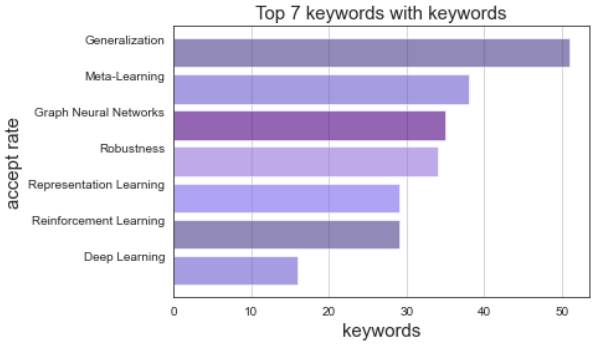

Fig. 6. Histogram of the relationship between top7 keywords and accept rate in 2021 (%)

The highest acceptance rate for 'Generalization' is 51%, and the lowest is 16% for 'Deep Learning'. The frequency of keywords is different from the frequency of keywords, confirming the previous conjecture that the frequency of keywords does not represent their acceptance rate. It shows that the number of articles with 'Generalization' as keywords in 2021 is low, but the acceptance rate is high. It can also be found from this graph that the more detailed the keywords are, the more likely they are to be accepted.

4) About rating and accept rate.

Based on the data processing in the previous section, data from 2021 were selected for the study. The histogram was plotted by combining decision and rating.

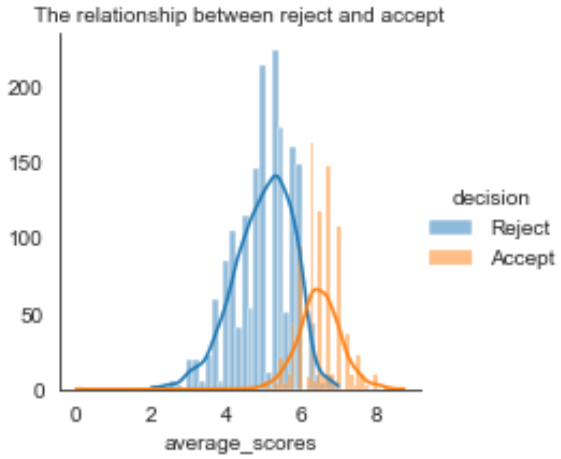

Fig. 7. Histogram of the relationship between average_scores and accept&reject in 2021

The figure shows that the median of average_scores for reject is roughly 5, and the median of accept is approximately 6. It offers a high probability of reject when the expert review score is below 5 and an increased likelihood of accept when it is above 6. The relationship between rating and accept can be roughly seen.

4.2 Affiliation data related

1) About organization.

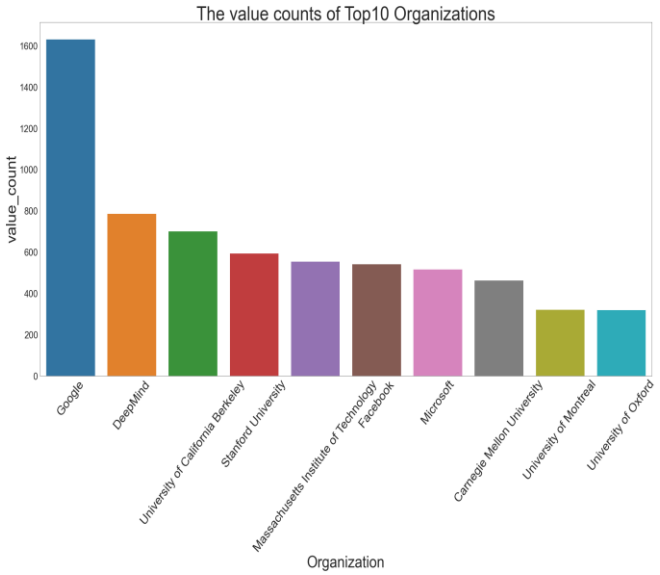

Fig. 8. Histogram of the frequency of organization from 2017-2021

Based on the data processing in the previous section, the top 25 organizations were selected for the five years from 2017 to 2021. The top25 organizations were selected and the top 25 organizations are mainly universities and Internet companies. As an American multinational technology company, Google has a significant advantage in areas such as artificial intelligence. Again, since the acceptance rate is not combined, it is impossible to say which organization is more likely to be accepted by ICLR because of its high submission frequency.

2) About organization and accept rate.

This part of the data was selected from the top10 organizations that appeared in 2021. taking 'Google' in the first place as an example, first of all, we found the mailbox of 2021paperlist that contains the domain name of Google's mailbox The first one is 'Google', and the first one is 'gmail', and the acceptance rate is calculated (the same treatment is done for other Internet companies). It is worth noting that due to the complexity of the school's email domain, this paper chose 'edu' as the suffix of the school's email. Still, statistics are not limited to the top ten organizations in terms of frequency, so there may be errors in the subsequent analysis.

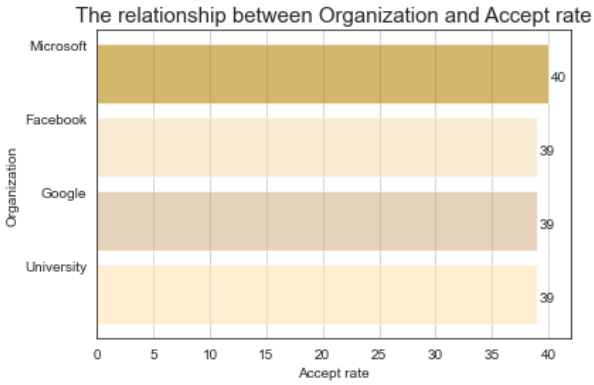

Fig. 9. Histogram of the relationship between top10 organizations and accept rate in 2021 (%)

Finally, plotted a histogram of organization and acceptance rate, even if all school data are selected, the acceptance rate of 'Microsoft' still ranks the highest at 40%, and the school at 39%. The school's acceptance rate in 2021 indicates that there is an abundance of talent for the future development of society.

5 Conclusion

5.1 Summary of findings

The topic of this paper is to use VDA to analyze the impact factor of a scientific paper accepted by ICLR, using the data of papers from ICLR for 2017-2021, and finding that the impact factors of papers accepted are keywords, rating and organization. it is worth noting that accept is not related to keywords and organization are not associated with the frequency of occurrence. Among them, rating reflects the influence of the expert review system on accept. keywords affect accept: whether keywords are detailed and specific, which indicates the need to add more micro-research value in future papers. Organization affects accept is the professionalism of the organization and the organizational environment. Of course, the data from the universities therein also show us that there is ample backup talent in the future in areas such as deep learning and AI.

A good machine learning environment, a more concrete paper topic, and a higher expert review score are the essential factors for a scientific paper to be accepted by ICLR. This paper also provides a reference for the research direction of machine learning papers.

5.2 The limitations of the study

Firstly, the number of features is small due to the single data type, so the analysis is one-sided. Secondly, due to the limited knowledge of python cannot successfully make the model of predicting accept, and there is a limitation for dealing with the relevance of text-based data. Thirdly, the classification of accept can be more specific, it can be

divided and analyzed by rating, for example, Oral has the highest rating, the spotlight is the second, and the lowest is poster. such analysis will make the accuracy of prediction model increase. Finally, due to the problem of data processing, the email domain name of the school is complicated, and 'edu' is chosen as the suffix of the school email in this paper. Still, such statistics are not limited to the top ten organizations regarding frequency, so there may be analysis errors in the follow-up.

5.3 Future area of study

Based on this, if more data and model experiments follow, a machine model to predict whether ICLR will accept a paper can be successfully built, giving scholars a great help in transforming research results into academic papers. Likewise, such a model will help predict the future direction of machine learning research.

References

1. Artificial Intelligence. Google Scholar Metrics. (2020). Retrieved 20 May 2022, from https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence.
2. Gao, J. (2022). QBUS6860-Individual Assignment 2 [Ebook] (p. 2). The University ofsydney. Retrieved 20 May 2022, from https://canvas.sydney.edu.au/courses/40419/pages/assessment-information-and-resources.
3. Ghosh, A., Pande, N., Goel, R., Mujumdar, R., & Sistla, S. (2019). Conference Paper Acceptance Prediction Acceptometer. Rohangoel.com. Retrieved 20 May 2022, from https://rohangoel.com/Acceptometer/.
4. J. Joshi, D., Kulkarni, A., Pande, R., Kulkarni, I., Patil, S., & Saini, N. (2021). Conference Paper Acceptance Prediction: Using Machine Learning. Machine Learning and Information Processing, 143–152. Retrieved 20 May 2022, from https://link.springer.com/chapter/10.1007/978-981-33-4859-2_14.
5. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., & Schwartz, R. (2018). A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. arXiv.org. Retrieved 20 May 2022, from https://arxiv.org/abs/1804.09635.
6. Momen, S., & SkorikovSifat, M. (2020). Machine learning approach to predicting the acceptance of academic papers. ResearchGate. Retrieved 20 May 2022, from https://www.researchgate.net/publication/343783292_Machine_learning_approach_to_predicting_the_acceptance_of_academic_papers.
7. Shaikh, A. (2021). 7 steps to publishing in a scientific journal. Elsevier Connect. Retrieved 21 May 2022, from https://www.elsevier.com/connect/7-steps-to-publishing-in-a-scientific-journal.
8. Wang, W., Zhang, J., Zhou, F., Chen, P., & Wang, B. (2021). Paper acceptance prediction at the institutional level based on the combination of individual and network features. Scientometrics, 126(2), 1581-1597. From https://doi.org/10.1007/s11192-020-03813-x.