ISSI Newsletter

QUARTERLY E-NEWSLETTER OF THE INTERNATIONAL SOCIETY FOR SCIENTOMETRICS AND INFORMETRICS
ISSN 1998-5460

#63 / VOLUME 16 NUMBER 3
SEPTEMBER 2020

CONTENTS

GTM GOES VIRTUAL

THE 10TH ANNUAL GLOBAL TECHMINING CONFERENCE "MINING SCIENCE & TECHNOLOGY INFORMATION RESOURCES IN TUMULTUOUS TIMES"

NOVEMBER 11-13, 2020 — VIRTUAL EVENT

The VP Institute, together with the Beijing Institute of Technology, is pleased to announce the 10th Global TechMining Conference as a wholly virtual event, accessible to all members of our global community. With the explosion of text-analytics driven by the necessity of advancing understanding of COVID-19, we feel it is more important than ever for TechMining practitioners to share knowledge. Details and registration information can be found at www.gtmconference.org.

Tech mining is text-oriented analytics that aims to generate practical intelligence from Science, Technology & Innovation (ST&I) information

ISSI e-Newsletter (ISSN 1998-5460) is published by ISSI (http://www.issi-society.org/).
Contributors to the newsletter should contact the editorial board by e-mail.

- **Wolfgang Glänzel**, *Editor-in-Chief*: wolfgang.glanzel[at]kuleuven.be
- **Balázs Schlemmer**, *Managing Editor*: balazs.schlemmer[at]gmail.com
- **Sarah Heeffer**, *Assistant Editor*: sarah.heeffer[at]kuleuven.**be**
- **Judit Bar-Ilan**
- **Sujit Bhatta**charya: sujit_academic[at]yahoo.**com**
- **María B**ordons: mbordons[at]cchs.cs**ic.es**
- **Juan Gorr**aiz: juan.gorraiz[at]univie.ac.**at**
- **Jacqueli**ne Leta: jleta[at]bioqmed.ufrj.**br**
- **Olle P**ersson: olle.persson[at]soc.umu.**se**
- **Ronald Ro**usseau: ronald.rousseau[at]kuleuven.**be**
- **Dietmar W**olfram: dwolfram[at]uwm.edu

Accepted contributions are moderated by the board. Guidelines for contributors can be found at http://www.issi-society.org/editorial.html. Opinions expressed by contributors to the Newsletter do not necessarily reflect the official position of ISSI. Although all published material is expected to conform to ethical standards, no responsibility is assumed by ISSI and the Editorial Board for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material therein.

GTM2017 Panel Discussion on "IP Data in TechMining" with (L to R) Dr. Martin G. Moehrle (Institute of Project Management and Innovation-Universitat Bremen), Rich Corken (Head of Data science and IP analytics in the UK's Intellectual Property Office's Informatics Team), and Dr. Alan Marco (2014-2017 Chief Economist of the United States Patent & Trademark Office)

resources – i.e., "Big Data" in various forms. It uses bibliometric and text-mining software (e.g., Derwent Data Analyzer (DDA), VantagePoint) and many other analytical & visualization tools.

Tech mining supports decision making in ST&I policy & management – e.g., competitive technical intelligence, R&D management, research evaluation, and coping with the COVID-19 pandemic. Expanding and novel data bring both opportunities and challenges. Researchers, analysts, and decision makers need to track ST&I dynamics together with rapidly evolving economic & societal factors.

GTM2020 seeks to bring together ST&I analysts, software specialists, researchers, and managers to advance text-data-driven solutions.

We are excited to engage the opportunities a virtual conference platform offers for greater participation diversity and equity! With recordings of livestream video, pre-recorded video, Q&A and discussion boards, etc. accessible beyond the 3-day agenda, we envision an exceptionally thoughtful exchange of knowledge.

Conference topics include and, as we dynamically assemble this first online GTM, will expand upon:

- Maximizing the potential of traditional and novel data – e.g.:
  - → ST&I + Other data sources (e.g. web scraping, social media, full-text information)
  - → Data refining (e.g. enhancing accuracy, disambiguation, natural language processing)

- Advancing and integrating methods – e.g.:
  - → Sharper topic/entity extraction with tools such as word embedding & clustering.
  - → Topic evolutionary path identification and science mapping
  - → Informetrics (e.g. bibliometrics, scientometrics); scientific visualization
  - → Predictive analytics; forecasting emerging technologies

GTM2018 TechMining for Global Good Award (L to R) Dr. Alan Porter (conference co-chair), Dr. Bruna Fonseca (Centro de Desenvolvimento Tecnológico em Saúde (CDTS) Fundação Oswaldo Cruz) and Denise Chiavetta (conference co-chair)

- Innovative analyses translating to useful intelligence – e.g.:
  - → Text mining in practice (e.g. bioinformatics, precision medicine)
  - → Competitive Technical Intelligence (CTI)
  - → Digest and synthesis of R&D re: COVID-19 pandemic

The conference programme includes:

1. Panel sessions – present multiple perspectives on an issue facing the field of tech mining. An overall abstract should describe the general session topic and how two, three, or four presenters offers coordinated oral presentations. Panel sessions will be scheduled as livestream video and include moderated Q&A.

2. Oral presentations – address well-advanced research (up to 20 minutes). Oral presentations are scheduled as livestream video or pre-recorded video and will include moderated Q&A.

3. Power talks – deliver a concise description of your research in-process and its impacts (up to 10 minutes). Provide an extended abstract (up to 500 words). Power Talks are scheduled as pre-recorded video with Q&A board.

4. Novel online presentations – based on suggestions on alternative modes of sharing tech mining knowledge.

The working language of the conference is English. Accepted contributions scheduled for presentation will be invited to submit full papers for consideration in a Scientometrics special issue dedicated to the conference theme.

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

12–15 JULY 2021, LEUVEN (BELGIUM)

VIRTUAL EVENT

Conference Chairs:

Koenraad Debackere, KU Leuven
Wolfgang Glänzel, KU Leuven
Bart Thijs, KU Leuven
Tim Engels, UAntwerp

Programme Chairs:

Ronald Rousseau (Belgium)
Lin Zhang (China/Belgium)
Henk F. Moed (Italy)

BACKGROUND

ISSI2021 is part of a long series of biennial conferences launched in 1987 that provide an international forum for scientists, research managers and administrators, as well as information professionals to share research and debate the advancements of informetric and scientometric theory and applications. The conference is organised by KU Leuven in close collaboration with the University of Antwerp under the auspices of ISSI – the International Society for Informetrics and Scientometrics (http://www.issi-society.org/).

Due to the constraints imposed by the COVID-19 pandemic situation and the resulting risks for planning and organising a full-fledged conference with physical presence of attendees, the organisers decided to go virtual next year. This decision will warrant a smooth organisation without unpredictable events and unnecessary modifications and adjustments during the preparation process. We thank all potential participants for their understanding and support in advance and duly hope for a high participation. The details of the organisation can be found on our website https://www.issi2021.org/.

SCOPE

The goal of ISSI2021 is to bring together scholars and practitioners in the domain of informetrics, bibliometrics, scientometrics, webometrics and altmetrics to discuss new research directions, advanced methods and theories, and to highlight the best research

IMPORTANT DATES

Full Papers, RiP, Workshop/Tutorial paper submission deadline	15 Jan 21
Paper/Workshop/Tutorial notification of acceptance/rejection	26 Feb 21
Poster submission deadline	22 Feb 21
Poster notification of acceptance/rejection	22 Mar 21
Doctoral Forum submission deadline	01 Mar 21
Doctoral Forum result announcement	05 Apr 21
Final paper/poster submission (at least one author must register)	26 Apr 21
Early Bird registration	15 Mar – 15 May 21
Conference	12 Jul – 15 Jul 21

in this area. In order to achieve this goal, we ask researchers worldwide to submit original research manuscripts, particularly full papers, research-in-progress papers or posters, to propose and organise tutorials and workshops, with a special emphasis on the future of this area and on its interdisciplinary links with other fields.

All manuscripts should be submitted to the conference editorial manager system (ConfTool) at https://www.conftool.pro/issi2021/. Manuscript templates for full, research-in-progress and poster papers can be downloaded from the conference website.

CONFERENCE TOPICS

With this scope in mind, major conference topics of interest include, but not limited to:

- Informetric theory
- Methods and techniques
- Citation and co-citation analysis
- Research collaboration, mobility and internationalisation
- Knowledge dissemination, integration and interdisciplinarity
- Bibliometric indicators – presence and future
- Webometrics and altmetrics
- Science mapping and visualization
- Science policy and research assessment
- University policy and institutional rankings
- Communication channels: periodicals, proceedings, books and electronic publications
- Knowledge discovery and data mining
- Bibliometrics-aided information retrieval
- Data sources and data processing
- Data harmonisation and integration
- Macro-, meso- and micro-level studies
- Open science – open access and open data
- Patent analysis
- Science-technology interface
- Data accuracy and disambiguation
- Scientific fraud and dishonesty

CONTACT

To get an answer to any conference-related question, the easiest way is contacting the Conference office via email: info@issi2021.org.

IMPORTANT NOTE

KU Leuven pursues a policy of equal opportunity and diversity, deplores and prosecutes any kind of discriminatory or otherwise unacceptable behaviour. All participants are encouraged to keep these principles in mind when registering for the conference. Further information on these principles can be found on the university website, cf. integrity and behaviour.

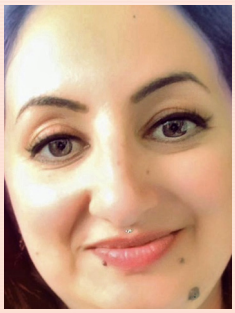
DOING SCIENCE IN TIMES OF CRISIS: SCIENCE STUDIES PERSPECTIVES ON COVID-19 (1st & 2nd EDITION)

WEBINAR REPORT

GRISCHA FRAUMANN

GIOVANNI COLAVIZZA

LUDO WALTMAN



ZOHREH ZAHEDI

INTRODUCTION

The webinar series "Doing science in times of crisis: Science studies perspectives on COVID-19" started on 20 May during early stages of the pandemic. The goal of this event is to connect and showcase COVID-19 science studies research from institutions around the world. At the same time, the webinar is also embedded into a broader research line on COVID-19 at CWTS, Leiden University. More information on the relation between the research line and the webinar series is available in an interview by Rodrigo Costas at the InSySPo São Paulo Excellence Chair.

THE 1st EDITION (20 MAY)

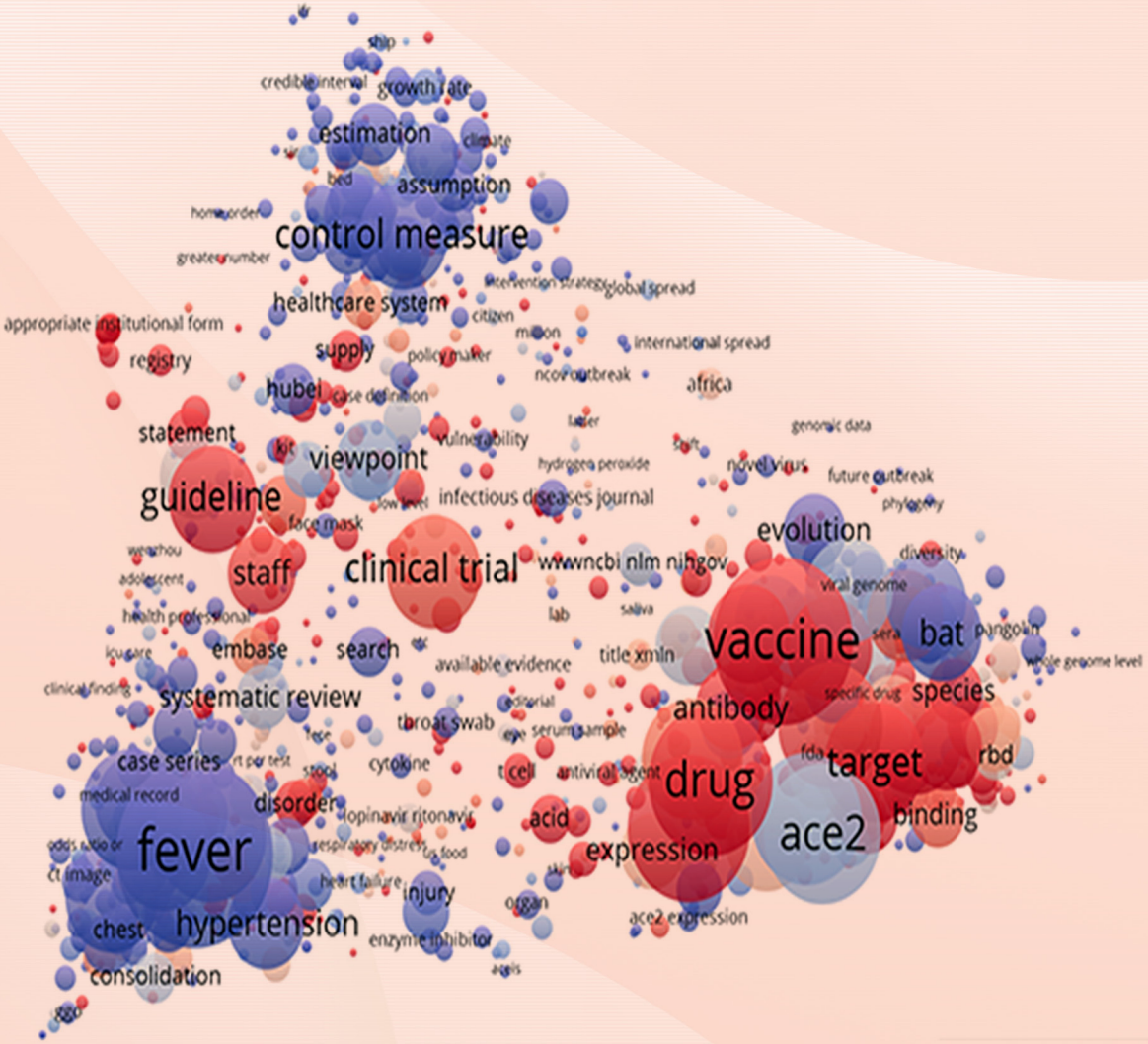
The 1st edition of the webinar was organized by Ludo Waltman and Giovanni Colavizza at CWTS, while the speakers were affiliated with other institutions. The webinar was attended by almost 200 participants. The webinar was organized into the following

three panels with a moderated discussion at the end of each panel.

Panel 1 "Debates on social media" included presentations on "COVID-19 publications: Citation indexes and altmetrics" by **Mike Thelwall** (University of Wolverhampton); "Media coverage of COVID-19 research" by **Mike Taylor** (Digital Science); and "Assessing the risks of "infodemics" in response to COVID-19 epidemics" by **Riccardo Gallotti** (Bruno Kessler Foundation – FBK, Trento).

Panel 2 „Societal questions" provided a rich perspective through the following presentations. "How scientific research reacts to international public health emergencies: a global analysis of response patterns" by **Lin Zhang** (Wuhan University); "Early signals of a widening gender gap in publication frequency during the COVID-19 pandemic" by **Jens Peter Andersen** (Aarhus University); and "Consolidation in a crisis: patterns of international collaboration in COVID-19 research" by **Caroline Wagner** (Ohio State University).

Finally, Panel 3, entitled „Mapping COVID-19 research" focused on the following talks. "The COVID-19 Open Research Dataset" was described by **Lucy Wang** and **Kyle Lo** (Semantic Scholar, Allen Institute for AI); "Tracking the growth of COVID-19 preprints" was presented by **Nicholas Fraser** (ZBW Leibniz Information Centre for Economics); and **Simon Porter** (Digital Science) dived deeper into "COVID-19 and preprint publishing culture".



Example of a visualization on clustered COVID-19 terms created with the VOSviewer software

Based on the positive feedback on the first edition, we started a 2nd edition, which was a coordinated effort by CWTS and TIB Leibniz Information Centre for Science and Technology. This edition was organized by Ludo Waltman, Zohreh Zahedi, Grischa Fraumann and Giovanni Colavizza.

THE 2nd EDITION (8 SEPTEMBER)

The feedback from the 1st edition was taken into account to develop the webinar series further, for example an additional focus was set on knowledge graphs, and we invited two speakers for that purpose. The 2nd edition took place on 8 September. The event was again structured into three panels with a moderated discussion at the end of each panel.

The speaker line up for Panel 1 „Pandemic effect" was as follows. **Milad Haghani** (University of Sydney) presented on "Covid-19 pandemic and the unprecedented mobilisation of scholarly efforts prompted by a health crisis: Scientometric comparisons across SARS, MERS and 2019-nCov literature"; **Wei Yang Tham** (Harvard University) focused on "Quantifying the Immediate Effects of the COVID-19 Pandemic on Scientists"; and **Serge Horbach** (Aarhus University) reported on "Pandemic Publishing: Medical journals strongly speed up their publication process for Covid-19". **Stefano Canali** and **Simon Lohse** (Leibniz University Hannover) presented together on "Epistemological aspects of evidence-based health policy: the case of Covid-19", and finally **Goran Murić** (University of Southern California) talked about "COVID-19 amplifies gender disparities in research".

Panel 2 „Social media" included the following two presentations. "COVID-19 research and social media" was introduced by **Rodrigo Costas** (Leiden University), and **Anders Blok** (University of Copenhagen) talked about "How We Tweet About Coronavirus, and Why: A Computational Anthropological Mapping of Political Attention on Danish Twitter during the COVID-19 Pandemic".

The final session, Panel 3 "Open science/open research" focused on a call for open science, and on the above-mentioned knowledge graphs. For that, **Jan Homolak** (University of Zagreb) presented on "Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders"; **Jennifer D'Souza** (TIB Leibniz Information Centre for Science and Technology) introduced "Covid-19 Bioassays in the Open Research Knowledge Graph"; and finally **Maria-Esther Vidal** (TIB Leibniz Information Centre for Science and Technology) described "How Do Knowledge Graphs Contribute to Understanding COVID-19 Related Treatments?".

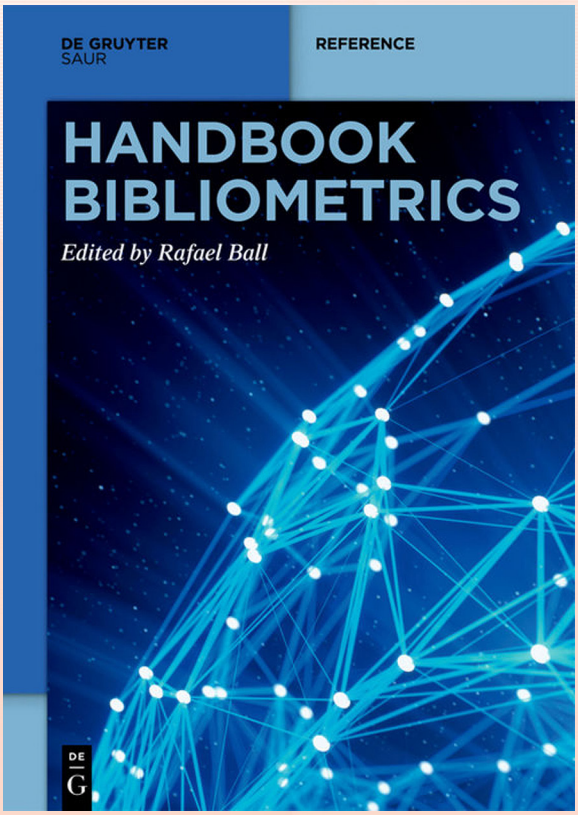
OUTLOOK

The webinar included diverse speakers and an audience from the science studies community and beyond. As in all current events, participation from around the world is possible due to the virtual format. We believe that the webinar provided an opportunity to connect COVID-19 science studies research and that the participants could benefit from the presentations and discussions. A 3 hours event structured into three panels seems to be an appropriate format for this purpose, although the different time zones of the organizers and participants are sometimes challenging.

Finally, the presentation videos of the 2nd edition will be edited by TIB Conference Recording Service and published at the TIB AV-Portal, a platform for scientific videos and events. We are currently discussing if a 3rd edition of the webinar series is on the horizon. Based on the positive feedback, we feel encouraged to do so. We would like to thank especially the engaged speakers who swiftly accepted our invitation to present at the webinars.

BOOK REVIEW: HANDBOOK BIBLIOMETRICS

RAFAEL BALL (ED.)
ETH-LIBRARY ZÜRICH, SWITZERLAND



A new Handbook on Bibliometrics will be published with De Gruyter soon.

OVERVIEW

"Bibliometrics and altmetrics are increasingly becoming the focus of interest in the context of research evaluation. The Handbook Bibliometrics provides a comprehensive introduction to quantifying scientific output in addition to a historical derivation, individual indicators, institutions, application perspectives and data bases. Furthermore, application scenarios, training and qualification on bibliometrics and their implications are considered."

Including the editorial introduction, the book comprises 45 chapters, which are organized in eight major parts reflecting the state-of the-art of contemporary bibliometric research:

- History and Institutionalization of Bibliometrics
- Theory, Principles and Methods of Bibliometrics
- (Classical) Indicators
- Alternative Metrics (Altmetrics)
- Applications, Practice and Special Issues in Bibliometrics
- The Data Basis in Bibliometrics
- Teaching and Training
- The Future of Bibliometrics

The handbook will be published as part of the De Gruyter Reference book series in December 2020. The volume will be available in print and as e-book in EPUB and PDF format.

INDIVIDUAL DOCUMENT CLASSIFICATION; PROMISING RESULTS FROM CONVOLUTIONAL NEURAL NETWORKS AND GRAPH EMBEDDINGS

BART THIJS

KU Leuven, ECOOM & Dept MSI,
Leuven, Belgium

bart.thijs@kuleuven.be

INTRODUCTION

Classification schemes used in most of the commercial multidisciplinary bibliographic databases are journal-based systems and lacking a proper document-based classification system. The WoS Core Collection uses their 'Web of Science Categories' which comprises about 250 subject areas and each journal indexed in this database is assigned to one or more subject category. In addition, Clarivate Analytics is applying their system of 22 research areas for the calculation of the Essential Science Indicators. This, too, is a journal-based system with each journal have a single assignment. Analogously, Scopus has a two-level journal classification scheme 'All Science Journal Classification' with 304 categories at the most fine-grained level and 27 top level categories. Several attempts have been conducted in the past to circumvent these shortcomings. Recently, Dimensions was released as the first bibliographic database with a document-based classification scheme but not without questions on its reliability and validity (Bornmann, 2020 or Singh et al., 2020). The topic has also been quite popular in some AI contests or hackathons. In this context, convolutional neural networks are often proposed for this task

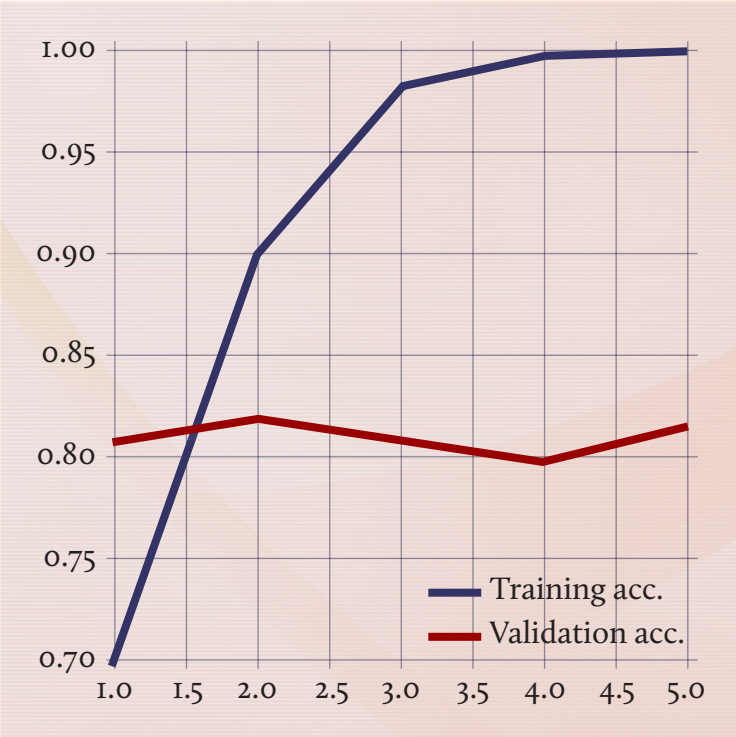


Figure 1: Accuracy of the Model 1 predictions

as CNN have a proven track record for text classification in other applications. Other approaches for the classification of individual scientific papers are based on citation links (Subelj et al, 2016; Waltman & van Eck, 2012 or Glänzel et al., 1999). In fact, this study tries to identify the added value of network data for the classification task as is done with hybrid approaches for document clustering and unsupervised learning (Thijs & Glänzel, 2018). The study does not present a ready-to-use article-based classification but tries to identify possible opportunities offered by the advent of new graph-based techniques and attempts to shed some light on possible shortcomings and hindrances that affects the reliability, validity and applicability. Important to mention here is the absence of a proper ground truth of the document classification.

DATA

Two publications sets have been used. First, a set of 40.790 publications indexed in the Web of Science Core Collection between 2007 and 2019 and assigned to a set of ISI subject categories compromising the field 'Non-internal medicine' of the Leuven-Budapest classification scheme and attributed to one of the nine disciplines within this field. The selected papers have between 3 and 7 citation links to other papers in the same field. The direction of the citation is neglected in this study. The construction of the datasets augments the probability that the paper is indeed properly assigned to the provided class. Consequently, the obtained classification models will not be applicable to a broad set of papers without additional training or adaptation. A second data set, mainly used for validation purposes, is a set of publications published in multidisciplinary journals during the same time period and citing or being cited by documents from the first set.

METHODS

Three classification models are compared in the study. First, two deep learning models, built around the combination of a convolutional neural network and a pooling layer preceded by a word embedding, are created within the Keras framework. The first one is a simple model with only one combination while the second one is more complex with three concatenated CNN's with each different filter size. The models are trained on a random selection of 70% of the first paper set. Both models produce for each input document an output vector with a predicted score for each of the 9 possible labels.

The third model is based on the StellarGraph implementation of the GraphSage framework (Hamilton, 2017) which creates low level node embeddings using not only the features of the node itself but also the features and labels of the neighborhood by applying a random walk selection procedure.

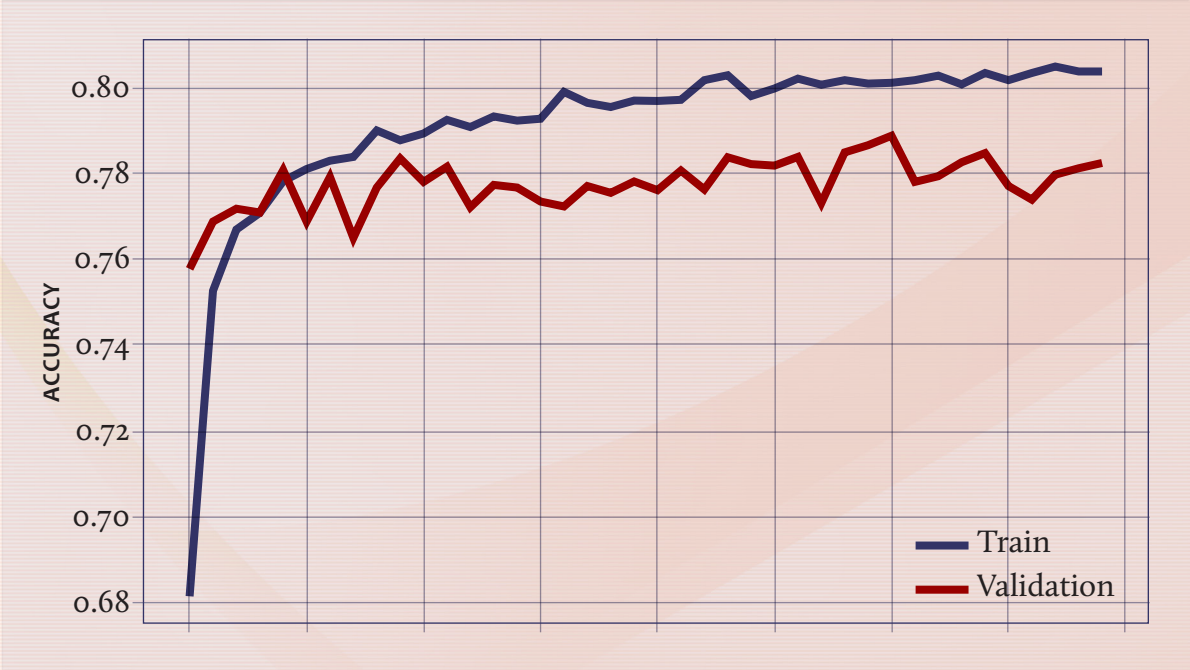


Figure 2: Accuracy of the Model 3 predictions

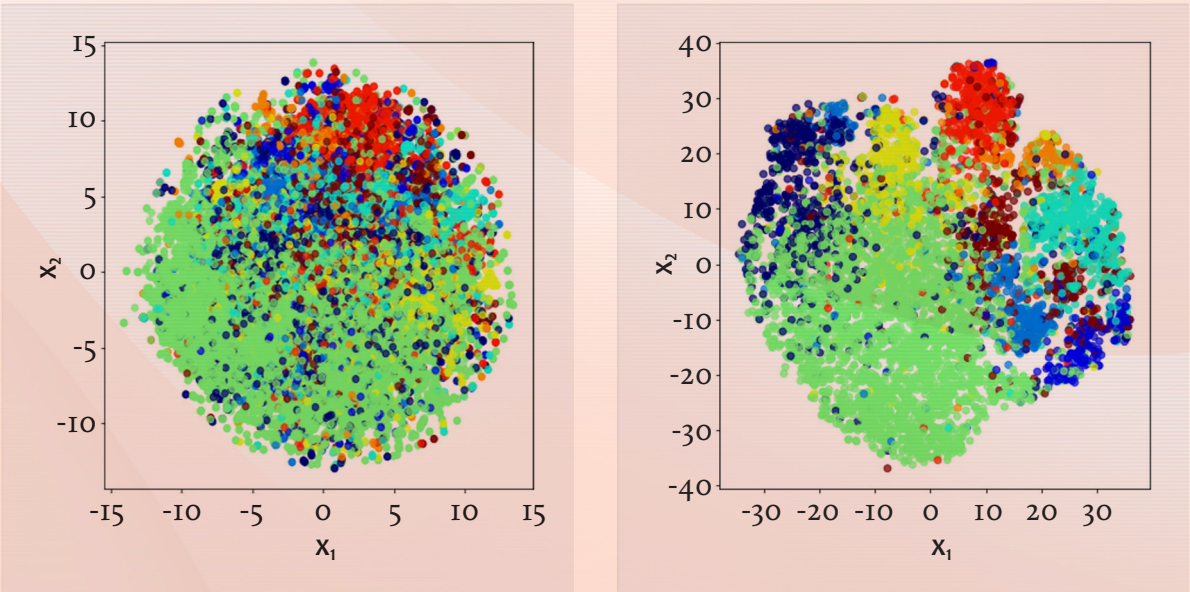


Figure 3: TSNE plot of Doc2Vec (left) and GraphSage (right) embeddings

An additional advantage of this framework is that it allows to efficiently generate representations of unseen data. The feature vector for the graph embedding is the result of a document embedding using Doc2Vec from the Gensim library with a dimensionality of 150.

RESULTS

The first model is trained in 5 epochs and reaches an accuracy of 99.95% for the training set being an overfitted result as the validation set has an accuracy of 81.33%. Learning the model took 2h and 20 minutes. Figure 1. plots the accuracy of the predictions on the training and validation set for the first model.

The second, more complex model took 27 hours and 23 minutes to reach a similar overfitted result for the training set and an accuracy of 82.97% for the training set. As the model is much slower in reaching its best results, the training took 25 epochs.

The third model starts from a Doc2Vec paragraph embedding and is using a random walk neighbor selecting mechanism

with up to 10 first order neighbors and up to 5 in the second order. The accuracy of the training set is 80.34% and for the validation set 78.23%. This shows clearly absence of an overfitting problem. The results are plotted in figure 2. This result was obtained after only 6 minutes with 40 epochs.

In order to have a fair comparison between the time needed for the training, the doc2vec embedding has to be taken into account. But also, being less the 40 minutes the total time from text and network data to prediction is well below one hour.

In fact, a TSNE plot (Figure 3) of both the document embedding using Doc2Vec and the node embedding from GraphSage shows the improvement of the classification adding the network information, the labels of citing and cited documents.

In a last step, the classification models have been applied on a selection of publications from multidisciplinary journals and manually validated. The results are promising as prediction scores can be used and deviations between the three predictions can be used for highlighting deviating cases.

REFERENCES

Bornmann, L., (2020). Field classification of publications in Dimensions: a first case study testing its reliability and validity. *Scientometrics*, 117 (1), 637-640.

Glänzel, W., & Schubert, A. & Czerwon, H.J., (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, 44(3), 427-439.

Hamilton, W.L., Ying, R., Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. arXiv:1706.02216.

Singh, P., Piryani, R., Singh, V.K. & Pinto, D., (2020). Revisiting subject classification in academic databases: A comparison of the classification accuracy of Web of Science, Scopus & Dimensions. *Journal of Intelligent & Fuzzy Systems*. 39(2), 2471-2476.

Subelj, L, van Eck, N.J., Waltman, L. (2016). Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods, PLOS ONE, 11 (4), e0154404.

Thijs, B. & Glänzel, W. (2018). The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics". *Scientometrics*, 115 (1), 21-33.

Waltman, L., & van Eck, N.J., (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378-2392.