

# Automatic Measurement of Corporate Reputation for Retail Companies from Online Public Data on the Web

**Author(s)**

Sitorus, Marselo; Loke, Rob

**Publication date**

2021

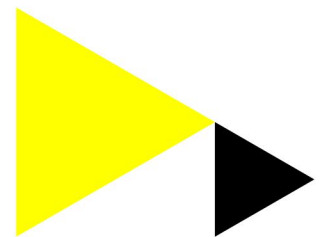
**Document Version**

Final published version

[Link to publication](#)

**Citation for published version (APA):**

Sitorus, M., & Loke, R. (2021). *Automatic Measurement of Corporate Reputation for Retail Companies from Online Public Data on the Web*. 34-35. Abstract from 10th International Conference on Data Science, Technology and Applications, Online .

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the library: <https://www.amsterdamuas.com/library/contact/questions>, or send a letter to: University Library (Library of the University of Amsterdam and Amsterdam University of Applied Sciences), Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

# DATA 2021

10<sup>th</sup> International Conference on Data Science, Technology  
and Applications

## Final Program and Book of Abstracts

6 - 8 July, 2021

<http://www.dataconference.org>

SPONSORED BY



PAPERS AVAILABLE AT



# DATA 2021

## Final Program and Book of Abstracts

---

10th International Conference on Data Science, Technology and  
Applications

Online Streaming  
July 6 - 8, 2021

**Sponsored by**

INSTICC - Institute for Systems and Technologies of Information, Control and Communication

**ACM In Cooperation**

ACM SIGMIS - ACM Special Interest Group on Management Information Systems



# Table of Contents

Foreword .....	5
Important Information .....	7
General Information .....	8
Program Layout .....	9
Tuesday Sessions: July 6 .....	17
Wednesday Sessions: July 7 .....	21
Thursday Sessions: July 8 .....	31
Author Index .....	39

## Foreword

This book contains the abstracts and final program of the 10th International Conference on Data Science, Technologies and Applications (DATA 2021) which is sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC), and held in cooperation with the ACM Special Interest Group on Management Information Systems (ACM SIGMIS). This year DATA was held as a web-based event due to the COVID-19 pandemic, from 6 - 8 July.

This conference brings together researchers, engineers and practitioners interested on databases, big data, data mining, data management, data security and other aspects of information systems and technology involving advanced applications of data.

The high quality of the DATA 2021 program is enhanced by the four keynote lectures, delivered by distinguished speakers who are renowned experts in their fields: Jan Recker (University of Cologne, Germany), Sandro Bimonte (INRAE, France), Hala Skaf-Molli (Nantes University, France), and Volker Markl (German Research Center for Artificial Intelligence (DFKI) and Technische Universität Berlin (TU Berlin), Germany).

DATA 2021 received 64 paper submissions from 27 countries of which 19% were accepted as full papers. In order to evaluate each submission, a double-blind paper review was performed by the Program Committee. All presented papers will be available at the SCITEPRESS Digital Library and will be submitted for indexation by Scopus, Google Scholar, DBLP, Semantic Scholar, Microsoft Academic, EI (Elsevier Index), and Web of Science (Clarivate). As in previous editions of the Conference, based on the reviewer's evaluations and the presentations, selected authors from the conference will be invited to submit extended versions of their papers for a book that will be published by Springer with the best papers of DATA 2021. This year, a short list of best papers will be invited for a post-conference special issue in the Springer Nature Computer Science Journal.

The program for this conference required the dedicated effort of many people. Firstly, we must thank the authors, whose research efforts are herewith recorded. Next, we thank the members of the Program Committee and the auxiliary reviewers for their diligent and professional reviewing. We would also like to deeply thank the invited speakers for their invaluable contribution and for taking the time to prepare their talks. Finally, we gratefully acknowledge the professional support of the conference secretariat and INSTICC team for all organizational processes, especially given the need to introduce online streaming, forum management, direct messaging facilitation and other web-based activities in order to make it possible for DATA 2021 authors to present their work and share ideas with colleagues in spite of the logistic difficulties caused by the current pandemic situation.

We wish you all an exciting conference and we look forward to having additional research results presented at the next edition of DATA.

Christoph Quix, Hochschule Niederrhein, University of Applied Sciences and Fraunhofer FIT, Germany  
Slimane Hammoudi, ESEO, ERIS, France  
Wil van der Aalst, RWTH Aachen University, Germany

# Important Information

## Event App

Download the Event App from the Play Store and App Store now, to have mobile access to the technical program and also to get notifications and reminders concerning your favorite sessions.

## Create Your Own Schedule \*

The option “My Program” gives you the possibility of creating a selection of the sessions that you plan to attend. This service also allows you to print-to-pdf all papers featured in your selection thus creating a pdf file per conference day.

## Online Access to the Proceedings \*

In the option “Proceedings and Final Program” you cannot only download the proceedings but also access the digital version of the book of abstracts with the final program.

## Digital Access to the Receipt \*

By clicking on the option “Delegate Home” and then “Registration Documents” it will enable you to access the final receipt which confirms the registration payment.

## Keynotes Videos

The keynote lectures will also be available on video on the website after the event, as long as the appropriate authorization from the keynote is received, so you will be able to see them again or watch them should you have missed one.

## Survey

Every year we conduct a survey to access the participants' satisfaction with the conference and gather the suggestions. You will receive an e-mail after the event with the detailed information. Your contribution will be carefully analysed and a serious effort to react appropriately will be made.

\* Please login to PRIMORIS ([www.insticc.org/Primoris](http://www.insticc.org/Primoris)), select the role “Delegate” and the correct event.

If you have any doubt, we will be happy to help you at the Welcome Desk.

# General Information

**Welcome Desk**

Tuesday, July 6 – Open from 14:15 to 18:00

Wednesday, July 7 – Open from 08:45 to 18:00

Thursday, July 8 – Open from 09:30 to 13:30

**Opening Session**

Tuesday, July 6, at 14:30 in the Plenary 1 room.

**Closing Session**

Thursday, July 8, at 13:15 in the Plenary 1 room.

**Secretariat Contacts**

DATA Secretariat

Address: Avenida de S. Francisco Xavier, Lote 7 Cv. C

2900-616 Setúbal, Portugal

Tel.: +351 265 520 185

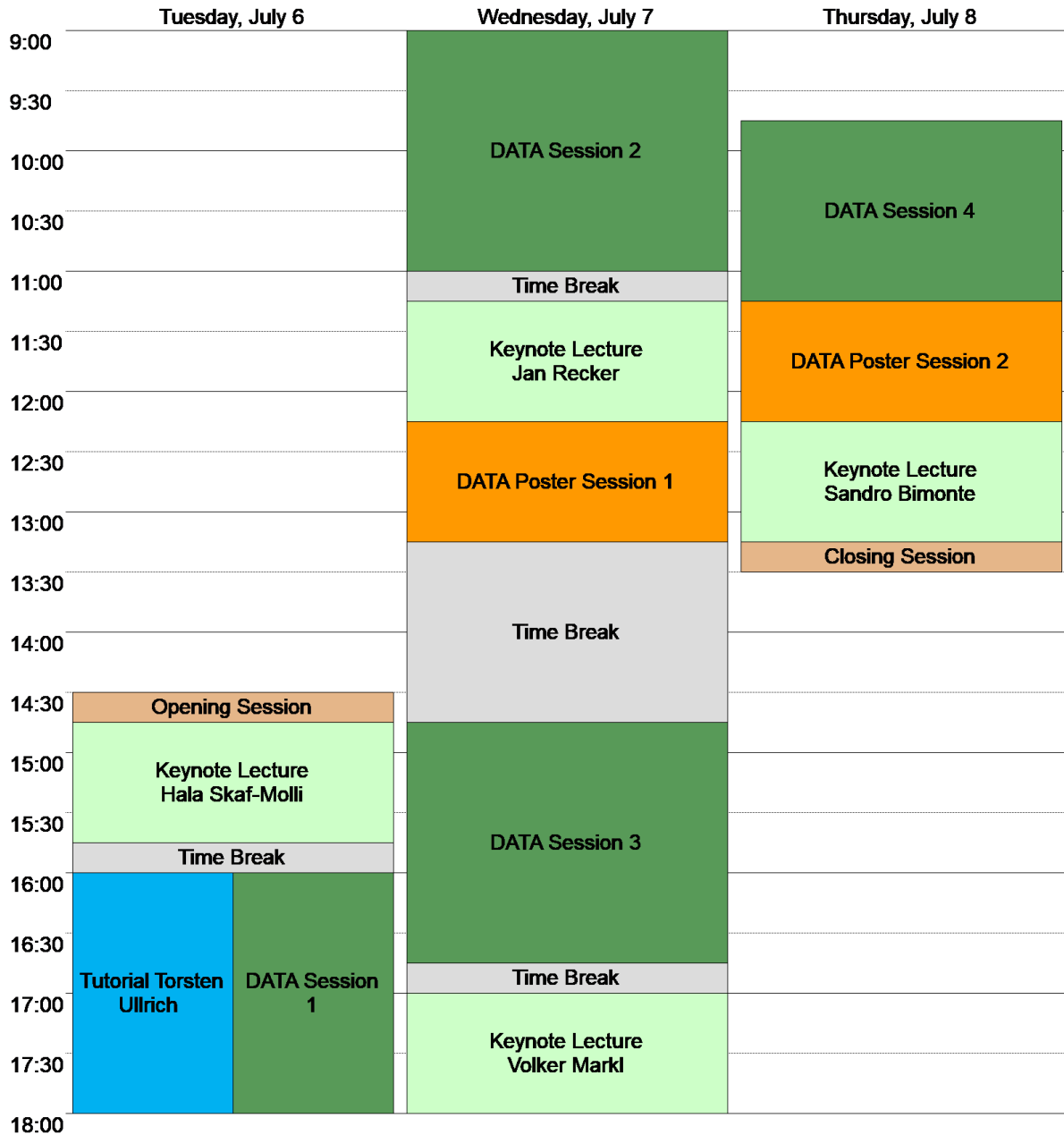
Fax: +351 265 520 186

e-mail: [data.secretariat@insticc.org](mailto:data.secretariat@insticc.org)

website: <http://www.dataconference.org>



# Program Layout



# **Final Program and Book of Abstracts**

---

# Contents

## Tuesday Sessions: July 6

### Opening Session (14:30 - 14:45)

Room Plenary 1 . . . . .	19
--------------------------	----

### Keynote Lecture (14:45 - 15:45)

Room Plenary 1 . . . . .	19
Querying Decentralized Knowledge Graphs, <i>by Hala Skaf-Molli</i> . . . . .	19

### Session 1 (16:00 - 18:00)

Room 2: <i>Data Science</i> . . . . .	19
<b>Complete Paper #31:</b> textPrep: A Text Preprocessing Toolkit for Topic Modeling on Social Media Data, <i>by Rob Churchill and Lisa Singh</i> . . . . .	19
<b>Complete Paper #60:</b> Forecasting Stock Market Trends using Deep Learning on Financial and Textual Data, <i>by Georgios-Markos Chatziloizos, Dimitrios Gunopulos and Konstantinos Konstantinou</i> . . . . .	19
<b>Complete Paper #14:</b> Deep Learning for RF-based Drone Detection and Identification using Welch's Method, <i>by Mahmoud Almasri</i> . . . . .	19
<b>Complete Paper #44:</b> Data Driven Hybrid Approach for Health Monitoring and Fault Detection in Military Ground Vehicles, <i>by Indu Shukla, Antoinette Silas, Haley Dozier, Brandon Hansen and W. Bond</i> . . . . .	20
<b>Complete Paper #50:</b> Estimating Territory Risk Relativity for Auto Insurance Rate Regulation using Generalized Linear Mixed Models, <i>by Shengkun Xie, Chong Gan and Clare Chua-Chow</i> . . . . .	20

### Tutorial (16:00 - 18:00)

Room 1 . . . . .	20
Introduction to the Data Analysis of Time Series, <i>by Torsten Ullrich</i> . . . . .	20

## Wednesday Sessions: July 7

### Session 2 (09:00 - 11:00)

Room 3: <i>Data Mining</i> . . . . .	23
<b>Complete Paper #11:</b> Similarity of Software Libraries: A Tag-based Classification Approach, <i>by Maximilian Auch, Maximilian Balluff, Peter Mandl and Christian Wolff</i> . . . . .	23
<b>Complete Paper #20:</b> A Comparative Study on Inflated and Dispersed Count Data, <i>by Monika Arora, Yash Kalyani and Shivam Shanker</i> . . . . .	23
<b>Complete Paper #6:</b> Data Mining for Animal Health to Improve Human Quality of Life: Insights from a University Veterinary Hospital, <i>by Oscar Tamburis, Elio Masciari, Christian Esposito and Gerardo Fatone</i> . . . . .	23
<b>Complete Paper #19:</b> A Survey of Social Emotion Prediction Methods, <i>by Abdullah Alsaedi, Phillip Brooker, Floriana Grasso and Stuart Thomason</i> . . . . .	23
<b>Complete Paper #23:</b> A Network based Approach for Reducing Variant Diversity in Production Planning and Control, <i>by Shailesh Tripathi, Sonja Strasser and Herbert Jodlbauer</i> . . . . .	23
Room 4: <i>Text Analytics</i> . . . . .	24
<b>Complete Paper #61:</b> A Graph-based Approach at Passage Level to Investigate the Cohesiveness of Documents, <i>by Ghulam Sarwar and Colm O'Riordan</i> . . . . .	24
<b>Complete Paper #62:</b> A Reference Process for Judging Reliability of Classification Results in Predictive Analytics, <i>by Simon Staudinger, Christoph Schuetz and Michael Schrefl</i> . . . . .	24
<b>Complete Paper #39:</b> Well-Being in Plastic Surgery: Deep Learning Reveals Patients' Evaluations, <i>by Joschka Kersting and Michaela Geierhos</i> . . . . .	24
<b>Complete Paper #8:</b> GRASP: Graph-based Mining of Scientific Papers, <i>by Navid Nobani, Mauro Pelucchi, Matteo Perico, Andrea Scrivanti and Alessandro Vaccarino</i> . . . . .	25
<b>Complete Paper #13:</b> A Comparison of Methods for the Evaluation of Text Summarization Techniques, <i>by Marcella Barbella, Michele Risi and Genoveffa Tortora</i> . . . . .	25

<b>Keynote Lecture (11:15 - 12:15)</b>	
<b>Room Plenary 1</b> . . . . .	25
From Representation to Mediation: Modeling Information Systems in a Digital World, <i>by Jan Recker</i> . . . . .	25
<b>Poster Session 1 (12:15 - 13:15)</b>	
<b>Room Posters DATA</b> . . . . .	25
<b>Abstract #13:</b> Real Estate Price Prediction with Artificial Intelligence Techniques, <i>by Sophia Zhou</i> . . . . .	25
<b>Abstract #18:</b> Knowledge Graph based Electrical Circuit Simulation and Component Selection, <i>by Rahman Syed, Johannes Bayer and Felix Thoma</i> . . . . .	26
<b>Complete Paper #2:</b> Determining How Different Factors Affect Police-Allegation's Sustainability in Chicago using Decision-Tree, <i>by Linxin Yang</i> . . . . .	26
<b>Complete Paper #3:</b> Archival and Museum Information as a Component of the Common Digital Space of Scientific Knowledge, <i>by N. Kalenov, I. Sobolevskaya and A. Sotnikov</i> . . . . .	26
<b>Complete Paper #4:</b> Building an Integrated Relational Database from Swiss Nutrition's (menuCH) and Multiple Swiss Health Datasets Acquired from 1992 to 2012 for Data Mining Purposes, <i>by Timo Lustenberger, Helena Jenzer and Farshideh Einsele</i> . . . . .	27
<b>Complete Paper #9:</b> Motif-based Classification using Enhanced Sub-Sequence-Based Dynamic Time Warping, <i>by Mohammed Alshehri, Frans Coenen and Keith Dures</i> . . . . .	27
<b>Complete Paper #21:</b> WFDU-net: A Workflow Notation for Sovereign Data Exchange, <i>by Heinrich Pettenpohl, Daniel Tebernum and Boris Otto</i> . . . . .	27
<b>Complete Paper #33:</b> Semantic Entanglement on Verb Negation, <i>by Yuto Kikuchi, Kazuo Hara and Ikumi Suzuki</i> . . . . .	27
<b>Complete Paper #38:</b> Using BPMN for ETL Conceptual Modelling: A Case Study, <i>by Bruno Oliveira, Óscar Oliveira and Orlando Belo</i> . . . . .	28
<b>Session 3 (14:45 - 16:45)</b>	
<b>Room 3: Data Management and Quality</b> . . . . .	28
<b>Complete Paper #7:</b> DERM: A Reference Model for Data Engineering, <i>by Daniel Tebernum, Marcel Altendeitering and Falk Howar</i> . . . . .	28
<b>Complete Paper #18:</b> DQ-MeeRKat: Automating Data Quality Monitoring with a Reference-Data-Profile-Annotated Knowledge Graph, <i>by Lisa Ehrlinger, Alexander Gindlhuber, Lisa-Marie Huber and Wolfram WöB</i> . . . . .	28
<b>Complete Paper #42:</b> Semantic Enrichment of Vital Sign Streams through Ontology-based Context Modeling using Linked Data Approach, <i>by Sachiko Lim, Rahim Rahmani and Paul Johannesson</i> . . . . .	28
<b>Room 4: Mobile Data and Data Integrity</b> . . . . .	29
<b>Complete Paper #30:</b> An Efficient Representation of Enriched Temporal Trajectories, <i>by Nieves Brisaboa, Antonio Fariña, Diego Otero-González and Tirso Rodeiro</i> . . . . .	29
<b>Complete Paper #27:</b> Database Recovery from Malicious Transactions: A Use of Provenance Information, <i>by Theppatorn Rhujittawiwat, John Ravan, Ahmed Saaudi, Shankar Banik and Csilla Farkas</i> . . . . .	29
<b>Complete Paper #36:</b> Invers Natural Number System to Maintain User-defined Sequence of Data Records, <i>by Seyfettin Öztürk</i> . . . . .	29
<b>Keynote Lecture (17:00 - 18:00)</b>	
<b>Room Plenary 1</b> . . . . .	29
Database Systems and Information Management: Trends and a Vision, <i>by Volker Markl</i> . . . . .	29
<b>Thursday Sessions: July 8</b>	
<b>Session 4 (09:45 - 11:15)</b>	
<b>Room 3: Data Science Applications</b> . . . . .	33
<b>Complete Paper #51:</b> Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques, <i>by Ahmad Homsí, Joyce Al Nemri, Nisma Naimat, Hamzeh Abdul Kareem, Mustafa Al-Fayoumi and Mohammad Abu Snober</i> . . . . .	33
<b>Complete Paper #12:</b> Biomedical Dataset Recommendation, <i>by Xu Wang, Frank van Harmelen and Zhisheng Huang</i> . . . . .	33
<b>Complete Paper #59:</b> Tailoring Taint Analysis for Database Applications in the K Framework, <i>by Md. Alam and Raju Halder</i> . . . . .	33
<b>Complete Paper #48:</b> Toward a Multimodal Multitask Model for Neurodegenerative Diseases Diagnosis and Progression Prediction, <i>by Sofia Lahrichi, Maryem Rhanoui, Mounia Mikram and Bouchra El Asri</i> . . . . .	33
<b>Room 8: Business Analytics</b> . . . . .	34
<b>Complete Paper #49:</b> A Study on the Effects of Response Time on Travel Package Attributes, <i>by Usha Ananthakumar and Sagun Pai</i> . . . . .	34
<b>Complete Paper #54:</b> A Longitudinal Model for Song Popularity Prediction, <i>by Ahmet Çimen and Enis Kayış</i> . . . . .	34
<b>Complete Paper #63:</b> A Company's Corporate Reputation through the Eyes of Employees Measured with Sentiment Analysis of Online Reviews, <i>by R. Loke and R. Lam-Lion</i> . . . . .	34

**Poster Session 2 (11:15 - 12:15)**

**Room Posters DATA** . . . . . 34

**Abstract #19:** Automatic Measurement of Corporate Reputation for Retail Companies from Online Public Data on the Web, by *Marselo Sitorus and Rob Loke* . . . . . 34

**Complete Paper #28:** Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks, by *Konstantinos Diamantaras, Michail Salampasis, Alkiviadis Katsalis and Konstantinos Christantonis* . . . . . 35

**Complete Paper #41:** Applied Feature-oriented Project Life Cycle Classification, by *Oliver Böhme and Tobias Meisen* . . . . . 35

**Complete Paper #45:** Impact of Duplicating Small Training Data on GANs, by *Yuki Eizuka, Kazuo Hara and Ikumi Suzuki* . . . . . 35

**Complete Paper #46:** Making Data Big for a Deep-learning Analysis: Aggregation of Public COVID-19 Datasets of Lung Computed Tomography Scans, by *Francesca Lizzi, Francesca Brero, Raffaella Cabini, Maria Fantacci, Stefano Piffer, Ian Postuma, Lisa Rinaldi and Alessandra Retico* . . . . . 36

**Complete Paper #52:** Knowledge Graph Analysis of Russian Trolls, by *Chih-yuan Li, Soon Chun and James Geller* . . . . . 36

**Complete Paper #53:** Aspect Based Sentiment Analysis on Online Review Data to Predict Corporate Reputation, by *R. Loke and W. Reitter* . . . . . 36

**Complete Paper #57:** Evo-Path: Querying Data Evolution through Complex Changes, by *Theodora Galani, Yannis Stavrakas, George Papastefanatos and Yannis Vassiliou* . . . . . 37

**Complete Paper #58:** Enhanced AI On-the-Edge 3D Vision Accelerated Point Cloud Spatial Computing Solution, by *Gaurav Kumar Wankar and Shubham Vohra* . . . . . 37

**Keynote Lecture (12:15 - 13:15)**

**Room Plenary 1** . . . . . 37

A Profile-aware Methodological Framework for Collaborative Multidimensional Modeling: Agro-biodiversity Case Study, by *Sandro Bimonte* . . . . . 37

**Closing Session (13:15 - 13:30)**

**Room Plenary 1** . . . . . 37



**Session 4B**  
**09:45 - 11:15**  
**Business Analytics**

**DATA**  
**Room 8**

Complete Paper #49

### A Study on the Effects of Response Time on Travel Package Attributes

Usha Ananthakumar and Sagun Pai  
*Indian Institute of Technology Bombay, Mumbai, India*

**Keywords:** Consumer Behavior, Conjoint Analysis, Demographic Profiling, Tourism Preferences, Willingness to Pay.

**Abstract:** The rapid growth of online surveys in the past decade has raised questions about the effects of response time on the results. The focus of our current study is to discuss the impact of response time on various travel package attributes, thereby understanding consumer cognitive process. This study makes use of a recently conducted conjoint analysis experiment on travel package preferences in order to gain insights into the impact of response time on attribute importance and willingness to pay (WTP). Accordingly, the respondents are grouped as fast and slow depending on their response time and their differences in conjoint attribute importance estimates are investigated. The study also examines the changes in consumer willingness to pay for the two groups. Additionally, the distinctions in socioeconomic characteristics between the fast and slow respondents are also analyzed. The results and conclusions obtained from this research will help tour operators to scrutinize the time taken by consumers and thereby deploy appropriate marketing strategy based on the respective importance values and WTP trends.

Complete Paper #54

### A Longitudinal Model for Song Popularity Prediction

Ahmet Çimen and Enis Kayış  
*Department of Industrial Engineering, Ozyegin University, Istanbul, Turkey*

**Keywords:** Music Analytics, Time-varying Coefficients, Mathematical Programming.

**Abstract:** Usage of new generation music streaming platforms such as Spotify and Apple Music has increased rapidly in the last years. Automatic prediction of a song's popularity is valuable for these firms which in turn translates into higher customer satisfaction. In this study, we develop and compare several statistical models to predict song popularity by using acoustic and artist-related features. We compare results from two countries to understand whether there are any cultural differences for popular songs. To compare the results, we use weekly charts and songs' acoustic features as data sources. In addition to acoustic features, we add acoustic similarity, genre, local popularity, song recentness features into the dataset. We applied Flexible Least Squares (FLS) method to estimate song streams and observe time-varying regression coefficients using a quadratic program. FLS method predicts the number of weekly streams of a song using the acoustic features and the additional features in the dataset while keeping weekly model differences as small as possible. Results show that the significant changes in the regression coefficients may reflect the changes in the music tastes of the countries.

Complete Paper #63

### A Company's Corporate Reputation through the Eyes of Employees Measured with Sentiment Analysis of Online Reviews

R. Loke and R. Lam-Lion  
*Centre for Market Insights, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands*

**Keywords:** Sentiment Analysis, Corporate Reputation, Natural Language Processing, Semantic Search, Scraping.

**Abstract:** Corporate reputation can be defined as the overall assessment of a company's performance over time (Kircova & Esen, 2018). Organizations with a positive corporate reputation create a competitive advantage and are more likely to influence customer's behaviors and attitudes (Kircova, 2018). Measuring corporate reputation from online data is an increasingly important area in business studies because the amount of opinions and comments is increasingly growing on the internet and has become very accessible to strangers (Shayaa, 2018). Traditionally, corporate reputation is measured with well-known approaches such as surveys, qualitative interviews, and sample groups (Smith, 2010). Researchers like Fombrun, Fonzy and Newburry (2015) developed instruments to measure corporate reputation and predictively modeled its impact on stakeholder outcomes. So far, however, there has been little attention in the literature on sophisticated measurement techniques for corporate reputation that can be applied to online reviews from the public web. This paper applies sentiment analysis in combination with semantic search as a suitable technique to explore how employees perceive organizations. By using our toolbox, organizations can adapt to market changes and cater to stakeholders' needs. Also, it can be used to raise awareness for organizations that are unaware of negative reviews online.

**Poster Session 2**  
**11:15 - 12:15**

**DATA**  
**Room Posters DATA**

Abstract #19

### Automatic Measurement of Corporate Reputation for Retail Companies from Online Public Data on the Web

Marselo Sitorus and Rob Loke  
*Centre for Market Insights, Amsterdam University of Applied Sciences, Amsterdam, The Netherlands*

**Keywords:** Aspect Based Sentiment Analysis (ABSA), Unsupervised Learning, Retail Industry, Corporate Reputation, Web Scraping, Online Reviews.

**Abstract:** Retail industry consists of the establishment of selling consumer goods (i.e. technology, pharmaceuticals, food and beverages, apparels and accessories, home improvement etc.) and services (i.e. specialty and movies) to customers through multiple channels of distribution including both the traditional brick-and-mortar and online retailing. Managing corporate reputation of retail companies is crucial as it has many advantages, for instance, it has been proven to impact generated revenues (Wang et al., 2016). But, in order to be able to manage corporate reputation, one has to be able to measure it, or, nowadays even better, listen to relevant social signals that are out there on the public web. One of the most extensive and widely used frameworks for measuring corporate reputation is through conducting elaborated surveys with respective stakeholders (Fombrun et al., 2015). This approach

is valuable but deemed to be laborious and resource-heavy and will not allow to generate automatic alerts and quick and live insights that are extremely needed in this era of internet. For these purposes a social listening approach is needed that can be tailored to online data such as consumer reviews as the main data source. Online review datasets are a form of electronic Word-of-Mouth (WOM) that, when a data source is picked that is relevant to retail, commonly contain relevant information about customers' perceptions regarding products (Pookulangara, 2011) and that are massively available.

The algorithm that we have built in our application provides retailers with reputation scores for all variables that are deemed to be relevant to retail in the model of Fombrun et al. (2015). Examples of such variables for products and services are high quality, good value, stands behind, and meets customer needs. We propose a new set of subvariables with which these variables can be operationalized for retail in particular. Scores are being calculated using proportions of positive opinion pairs such as <fast, delivery> or <rude, staff> that have been designed per variable. With these important insights extracted, companies can act accordingly and proceed to improve their corporate reputation. It is important to emphasize that, once the design is complete and implemented, all processing can be performed completely automatic and unsupervised.

The application makes use of a state of the art aspect-based sentiment analysis (ABSA) framework because of ABSA's ability to generate sentiment scores for all relevant variables and aspects. Since most online data is in open form and we deliberately want to avoid labelling any data by human experts, the unsupervised aspectator algorithm has been picked. It employs a lexicon to calculate sentiment scores and uses syntactic dependency paths to discover candidate aspects (Bancken et al., 2014).

We have applied our approach to a large number of online review datasets that we sampled from a list of 50 top global retailers according to National Retail Federation (2020), including both offline and online operation, and that we scraped from trustpilot, a public website that is well-known to retailers.

The algorithm has carefully been evaluated by manually annotating a randomly sampled subset of the datasets for validation purposes by two independent annotators. The Kappa's score on this subset was 80%.

best classifier achieves a predictive accuracy of almost 98%. This result is competitive with other state-of-the-art methods, which affirms that accurate and scalable purchasing intention prediction for e-commerce, using only session-based data, is feasible without any intense feature engineering.

Complete Paper #41

## Applied Feature-oriented Project Life Cycle Classification

Oliver Böhme and Tobias Meisen

*Chair for Technologies and Management of Digital Transformation,  
Bergische Universität Wuppertal, Rainer-Gruenter-Str. 21, Wuppertal,  
Germany*

**Keywords:** Machine Learning, Classification, Prediction, Deep Neural Networks, MLP, LSTM, Multivariate, Automotive, R&D, Projects Progressions, Project Life Cycle, Comparative Analysis.

**Abstract:** The increasing complexity in automotive product development is forcing traditional manufacturers to fundamentally rethink. As a result, many companies are already investing in the development of methods to increase the controllability of their development processes. The use of data-driven approaches is a promising way to provide an early prediction of potential problems in the course of a project by learning from the past. In vehicle development, projects can be divided into two basic categories: new vehicle launches and model enhancement projects. The course of projects according to the above-mentioned categories can be based on different influencing factors. To verify this hypothesis and to determine the extent of the differences in the data, we carry out a data-driven classification of the project category. In contrast to the recognition of other time-dependent data (e.g., univariate sensor data courses), we use multivariate project information from the automotive industry. With this paper, which is of an application nature, we prove that a multivariate classification of automotive projects can be realized based on the underlying project's progression.

Complete Paper #28

## Predicting Shopping Intent of e-Commerce Users using LSTM Recurrent Neural Networks

Konstantinos Diamantaras, Michail Salampasis, Alkiviadis Katsalis and Konstantinos Christantonis

*Intelligent Systems Laboratory, Department of Information and Electronic Engineering, International Hellenic University, Sindos, Thessaloniki, Greece*

**Keywords:** Purchase Intent, e-Commerce, LSTM-RNN, Web Usage Mining.

**Abstract:** An e-commerce web site is effective if it turns visitors into buyers achieving a high conversion rate. To this realm, it is useful to predict each user's purchase intent and understand their navigation behavior. Such predictions may be utilized to improve web design and to personalize shopper's experience, hopefully leading to increased conversion rates. Additionally, if such predictions can be done in real-time, during the ongoing navigation of an e-commerce user, the e-commerce application can take proactive stimuli actions to offer incentives with a view to increase the probability that a user will finally make a purchase. This paper presents a method for predicting in real-time the shopping intent of e-commerce users using LSTM recurrent neural networks. We test several variants of our method in a dataset created from the processing of Web server logs of an industry e-commerce web application, dividing user sessions in three different classes: browsing, cart abandonment, purchase. The

Complete Paper #45

## Impact of Duplicating Small Training Data on GANs

Yuki Eizuka<sup>1</sup>, Kazuo Hara<sup>1</sup> and Ikumi Suzuki<sup>2</sup>

<sup>1</sup> Yamagata University, 1-4-12 Kojirakawa-machi, Yamagata City, 990-8560, Japan

<sup>2</sup> Nagasaki University, 1-14 Bunkyo, Nagasaki City, 852-8521, Japan

**Keywords:** Generative Adversarial Networks, Small Training Data, Emoticons.

**Abstract:** Emoticons such as (‘\_’) are face-shaped symbol sequences that are used to express emotions in text. However, the number of emoticons is miniscule. To increase the number of emoticons, we created emoticons using SeqGANs, which are generative adversarial networks for generating sequences. However, the small number of emoticons means that few emoticons can be used as training data for SeqGANs. This is concerning because as SeqGANs underfit small training data, generating emoticons using SeqGANs is difficult. To address this problem, we duplicate the training data. We observed that emoticons can be generated when the duplication magnification is of an appropriate value. However, as a trade-off, it was also observed that SeqGANs overfit the training data, i.e., they produce emoticons that are exactly the same as the training data.