Peter Haber
Thomas Lampoltshammer
Manfred Mayr *Eds.*

Data Science – Analytics and Applications

Proceedings of the 2nd International Data Science Conference – iDSC2019

EBOOK INSIDE

Springer Vieweg

Data Science – Analytics and Applications

Peter Haber • Thomas Lampoltshammer
Manfred Mayr
Editors

Data Science – Analytics and Applications

Proceedings of the 2nd International Data Science Conference – iDSC2019

Springer Vieweg

Editors
Peter Haber
Informationstechnik & System-Management
Fachhochschule Salzburg
Puch/Salzburg, Österreich

Manfred Mayr
Informationstechnik & System-Management
Fachhochschule Salzburg
Puch/Salzburg, Österreich

Thomas Lampoltshammer
Dept. für E-Governance in Wirtschaft
und Verwaltung
Donau-Universität Krems
Krems an der Donau, Österreich

ISBN 978-3-658-27494-8 ISBN 978-3-658-27495-5 (eBook)
https://doi.org/10.1007/978-3-658-27495-5

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Springer Vieweg

Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Preface

It is with deep satisfaction that we write this foreword for the proceedings of the 2nd International Data Science Conference (iDSC) held in Salzburg, Austria, May 22nd - 24th 2019. The conference programme and the resulting proceedings represent the efforts of many people. We want to express our gratitude towards the members of our program committee as well as towards our external reviewers for their hard work during the reviewing process.

iDSC proofed itself – again – as an innovative conference, which gave its participants the opportunity to delve into state-of-the-art research and best practice in the fields of data science and data-driven business concepts. Our research track offered a series of presentations by researchers regarding their current work in the fields of data mining, machine learning, data management, and the entire spectrum of data science.

In our industry track, practitioners demonstrated showcases of data-driven business concepts and how they use data science to achieve organizational goals, with a focus on manufacturing, retail, and financial services. Within each of these areas, experts described their experience, demonstrated their practical solutions, and provided an outlook into the future of data science in the business domain.

Our sponsors – The MathWorks GmbH and Cognify KG – had their own, special platform via workshops to provide our participants hands-on interaction with tools or to learn approaches towards concrete solutions. In addition, an exhibition of products and services offered by our sponsors took place throughout the conference, with the opportunity for our participants to seek contact and advice.

Completing the picture of our programme, we proudly presented keynote presentations from leaders in data science and data-driven business, both researchers and practitioners. These keynotes provided all participants the opportunity to come together and share views on challenges and trends.

Keynote presentations were given by: Josef Waltl (Amazon Web Services), Bodo Hoppe (IBM Research & Development GmbH), Christian D. Blakely (PricewaterhouseCoopers Switzerland), Peter Parycek (Danube University Krems, ÖFIT/Fraunhofer Fokus Berlin), Stefan Wegenkittl (Salzburg University of Applied Sciences) and David C. Anastasiu (San José State University).

We thank all speakers for sharing their insights with our community. We also thank Michael Ruzicka for moderating the industry track and for his help and support for its realisation. Especially we like to thank our colleagues, Nicole Siebenhandl, Maximilian Tschuchnig and Dominik Vereno, for their enormous and contructive commitment to organizing and conducting the iDSC.

We are convinced that the iDSC proceedings will give scientists and practitioners an excellent reference to the current activities in the field of data science and also impulses for further studies, research activities and applications in all discussed areas. The visibility is ensured by the support of our publishing house Springer / Vieweg Wiesbaden Germany.

Peter Haber, Thomas Lampoltshammer and Manfred Mayr

Conference Chairs

Data Science is everywhere

The 2nd International Data Science Conference (iDSC2019) combined state-of-the-art methods from academia with industry best practices.

From 22nd to 24th May 2019, representatives of the academic and industry world met for the 2nd International Data Science Conference at Salzburg University of Applied Sciences, Austria. Over 40 top-class speakers from industry and academia presented trends and technologies, as well as new methods and analytical approaches; overall, more than 120 international experts took the opportunity to discuss new challenges and developments regarding data science in industry, research, education, and society.

The technological progress in IT and the resulting changes in the business and social environment ("digitalization") have led to an increase of the amount of data. Several billion terabyte of data have been produced up until now, and the number is growing exponentially. "Data collection, permeates all areas of our lives: from standard use cases such as the analysis of our shopping behaviour followed by purchase suggestions, up to competitive sports", explains conference organizer Peter Haber, senior lecturer at the degree programme Information Technology & Systems Management, who has initiated the conference together with Manfred Mayr (Salzburg University of Applied Sciences) and Thomas Lampoltshammer (Danube University Krems).

In many places, data mining and machine learning have become part of the occupational routine in order to generate value out of the flood of data. "The permeation in all areas was evident at the conference. The presented use cases ranged from market and trend analysis, the energy, health and sports sector, to predictive maintenance in the industrial sector", summarizes Manfred Mayr, co-organizer and academic programme director of the degree programme Business Informatics and Digital Transformation, the key areas of the conference.

Besides talks about state-of-the-art methods of data analytics by researchers from the international academic community, business experts gave insights into best practices and obstacles from the practice. The conference highlights:

- **Artificial intelligence and machine learning gain importance**

  More and more manufacturing companies place greater emphasis on machine learning, a form of artificial intelligence that enables systems to learn independently from data. They, for example, optimize their maintenance with this method. Monitors, among others, get equipped with intelligent sensors to collect data. For analysing the data, so-called predictive anticipation algorithms give information about the projected maintenance need. With this, a lot of money can be saved.

- **Developing the business models of the future with blockchain technology**

  Bodo Hoppe, distinguished engineer at IBM Research & Development GmbH, explained in his keynote that the potential for new business models lies in the continuous monitoring of the production and supply chain. With "IBM© Food Trust" the software giant IBM has developed a new trust and transparency system for the food industry. Hoppe: "The solution is based on blockchain technology and grants authorised users immediate access to meaningful data of the food supply chain, from the farm via the stores to the end consumer."

- **Small Data: Solutions for SME**

  The 2nd International Data Science Conference showed possibilities for the optimization of business models, especially for small and medium sized enterprises (SME). "Often they are lacking the special knowledge and the supposed data to evaluate the chances and risks for their own company", knows conference organizer Peter Haber. "Recent improvements in the area of deep learning, however, provide solutions for these cases with only limited data available", says David Anastasiu, assistant professor at the Department of Computer Engineering at San José State University. Thus, SME could improve their business operations significantly.

  Josef Waltl, Global Segment Lead at Amazon Web Services and alumnus of the degree programme Information Technology and Systems Management, presented, therefore, in his keynote a cloud-based machine-learning system that could help companies with their use cases.

- **Open data & scientific cooperation as chances for SME**

  Peter Parycek, head of the Department for E-Governance and Administration at Danube University Krems, wants to open the access to the data of the big players. SME could use these open data as a reference to generate transfer possibilities for their use cases. Experts, therefore, advice SME to invest in data scientists.

  Stefan Wegenkittl, academic programme director of the degree programme Applied Image & Signal Processing, as well as head of the applied mathematics and data mining department at the degree programme Information Technology and Systems Management, emphasized the possibility to intensify the exchange with higher education institutions: "Suitable solutions can be found if agile development and management processes are connected with current data science research questions."

- **Social benefits: detecting autism & revealing insider trade**

  Data science offers exciting possibilities not only for business purposes. David Anastasiu from the San José University presented his research project about the computer-based detection of autism. Anastasiu and his team received the respective data from electrocardiograms, which measure the activity of all heart muscle fibres, and data about the skin conductance.

  With this information, they determined the course of the relaxation times which – as they could show in their research project – can give an indication about whether a person is suffering from autism or not. Anastasiu: "Our model reached an accuracy of 99.33 percent. Diagnosis by doctors are around 82.9 percent accurate." Jad Rayes and Priya Mani from the George Mason University presented another field of application with social benefit. They are able to detect insider trade activities. With this information, the police could reduce the crime on the capital market.

Positive feedback for the conference

The high number of participants, as well as their positive feedback proved that the concept of the conference is successful. "The participants liked the format as well as the size. The great mix of academic inputs and practical examples from the industry was appreciated", so Haber proudly. The next conference will take place from 13th to 14th May 2020. "But this time in Dornbirn, in cooperation with FH Vorarlberg" – another indicator for the successful concept.

Data Science at Salzburg University of Applied Sciences

With a particular specialization in Data Science & Analytics in the master's programme, the degree programme Information Technology & Systems Management at Salzburg University of Applied Sciences offers prospective data scientists a perfect education. Alumni have, besides mathematical and statistical core competences, fundamental knowledge in the areas machine learning, data mining and deep learning. Data science is also one of the cross-cutting topics taught in the Business Informatics and Digital Transformation study degree.

Nicole Siebenhandl, Sandra Lagler
Local Organizer, Communications

Organization

Organizing Institution

Salzburg University of Applied Sciences

Conference Chairs

Peter Haber	Salzburg University of Applied Sciences
Thomas J. Lampoltshammer	Danube University Krems
Manfred Mayr	Salzburg University of Applied Sciences

Local Organizers

Nicole Siebenhandl	Salzburg University of Applied Sciences
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences
Dominik Vereno	Salzburg University of Applied Sciences

Program Committee

David Anastasiu	San José State University
Arne Bathke	University of Salzburg
Markus Breunig	University of Applied Science Rosenheim
Frank Danielsen	University of Agder
Eric Davis	Industrial Labs
Günther Eibl	Salzburg University of Applied Sciences
Süleyman Eken	Kocaeli University
Karl Entacher	Salzburg University of Applied Sciences
Mohammad Ghoniem	Luxembourg Institute of Science and Technology
Elmar Kiesling	Vienna University of Technology
Michael Gadermayr	Salzburg University of Applied Sciences
Manolis Koubarakis	National and Kapodistrian University of Athens
Maria Leitner	Austrian Institute of Technology
Elena Lloret Pastor	University of Alicante
Giuseppe Manco	University of Calabria
Robert Merz	Vorarlberg University of Applied Sciences
Edison Pignaton de Freitas	Federal University of Rio Grande do Sul
Florina Piroi	Vienna University of Technology
Kathrin Plankensteiner	Vorarlberg University of Applied Sciences

Siegfried Reich	Salzburg Research
Peter Reiter	Vorarlberg University of Applied Sciences
Michael Ruzicka	Cockpit-Consulting
Marta Sabou	Vienna University of Technology
Johannes Scholz	Graz University of Technology, Institute of Geodesy
Axel Straschil	pmIT Consult
Lőrinc Thurnay	Danube University Krems
Andreas Unterweger	Salzburg University of Applied Sciences
Gabriela Viale Pereira	Danube University Krems
Stefan Wegenkittl	Salzburg University of Applied Sciences
Karl-Heinz Weidmann	Vorarlberg University of Applied Sciences
Anneke Zuiderwijk	van Eijk – Delft University of Technology

Reviewer

David Anastasiu	San José State University
Frank Danielsen	University of Agder
Eric Davis	Industrial Labs
Günther Eibl	Salzburg University of Applied Sciences
Süleyman Eken	Kocaeli University
Karl Entacher	Salzburg University of Applied Sciences
Michael Gadermayr	Salzburg University of Applied Sciences
Peter Haber	Salzburg University of Applied Sciences
Manolis Koubarakis	National and Kapodistrian University of Athens
Thomas J. Lampoltshammer	Danube University Krems
Maria Leitner	Austrian Institute of Technology
Robert Merz	Vorarlberg University of Applied Sciences
Elena Lloret Pastor	University of Alicante
Edison Pignaton de Freitas	Federal University of Rio Grande do Sul
Florina Piroi	Vienna University of Technology
Kathrin Plankensteiner	Vorarlberg University of Applied Sciences
Siegfried Reich	Salzburg Research
Peter Reiter	Vorarlberg University of Applied Sciences
Johannes Scholz	Graz University of Technology, Institute of Geodesy
Lőrinc Thurnay	Danube University Krems
Maximilian E. Tschuchnig	Salzburg University of Applied Sciences
Andreas Unterweger	Salzburg University of Applied Sciences
Gabriela Viale Pereira	Danube University Krems
Stefan Wegenkittl	Salzburg University of Applied Sciences

Table of Content

German Abstracts

Data Analytics | Complexity

Exploring Insider Trading Within Hypernetworks
Erforschung von Insiderhandel innerhalb von Hypernetzwerken

Insiderhandel kann lähmende Auswirkungen auf die Wirtschaft haben, und seine Verhinderung ist entscheidend für die Sicherheit und Stabilität der globalen Märkte. Es wird angenommen, dass Insider, die zu ähnlichen Zeiten handeln, Informationen austauschen. Wir analysieren 400 Unternehmen und 2.000 Insider und identifizieren interessante Handelsmuster in diesen Netzwerken, die auf illegale Aktivitäten hinweisen können. Insider werden entweder als routinemäßige oder opportunistische Händler eingestuft, sodass wir uns auf gut getaktete und hochprofitable Handelsaktivitäten des letzteren Typs konzentrieren können. Durch die Einstufung des Handels und der Analyse der Rolle jedes Händlers in einem Hypernetzwerk zeigen sich Gruppen von opportunistischen und routinemäßigen Händlern. Diese Idee bildet die Grundlage eines graphenbasierten Erkennungsalgorithmus, der darauf abzielt, Händler zu identifizieren, die zu opportunistischen Gruppen gehören. Die Handelseinstufung und Handelsgruppen bieten interessante Möglichkeiten, zuverlässigere Überwachungssysteme zu entwickeln, die automatisch illegale Aktivitäten auf den Märkten erkennen und vorhersagen können, mit welcher Wahrscheinlichkeit diese Aktivitäten in Zukunft eintreten werden.

Chance influence in datasets with a large number of features
Zufallseinfluss bei Datensätzen mit einer großen Anzahl von Merkmalen

Die maschinelle Lernforschung, z. B. Genomforschung, basiert oft auf wenigen Datensätzen, die zwar sehr viele Merkmale, aber nur kleine Stichprobengrößen enthalten. Diese Konfiguration fördert den Zufallseinfluss auf den Lernprozess und die Auswertung. Frühere Forschungen konzentrierten sich auf die Verallgemeinerung von Modellen, die auf Grundlage solcher Daten erhalten wurden. In diesem Beitrag untersuchen wir den Zufallseinfluss auf die Klassifizierung und Regression. Wir zeigen empirisch, wie groß der Zufallseinfluss auf diese Datensätze ist. Damit werden die daraus gezogenen Schlussfolgerungen in Frage gestellt. Wir verknüpfen die Beobachtungen der Zufallskorrelation mit dem Problem der Methodengeneralisierung. Schließlich besprechen wir die Zufallskorrelation und nennen Richtlinien, die den Zufallseinfluss verringern.

Data Analytics | NLP and Semantics

Combining Lexical and Semantic Similarity Methods for News Article Matching
Kombination von lexikalischen und semantischen Ähnlichkeitsmethoden beim Abgleich von Nachrichtenartikeln

Der Abgleich von Nachrichtenartikeln verschiedener Quellen mit unterschiedlichen Schilderungen ist ein entscheidender Schritt für die verbesserte Verarbeitung des Online-Nachrichtenflusses. Obwohl es Studien zum Auffinden gleicher oder nahezu gleicher Dokumente in verschiedenen Bereichen gibt, beschäftigt sich keine dieser Studien mit der Gruppierung von Nachrichtentexten auf Grundlage ihrer Ereignisse oder Quellen. Ein bestimmtes Ereignis kann aufgrund der unterschiedlichen politischen Ansichten der Verlage aus sehr unterschiedlichen Perspektiven mit unterschiedlichen Wörtern, Konzepten und Meinungen geschildert werden. Wir entwickeln eine neue Methode zum Abgleich von Nachrichtendokumenten, die mehrere verschiedene lexikalische Abgleichswerte mit Ähnlichkeitswerten aufgrund von semantischen Darstellungen von Dokumenten und Wörtern kombiniert.

Unser experimentelles Ergebnis zeigt, dass diese Methode beim Nachrichtenabgleich sehr erfolgreich ist. Wir entwickeln darüber hinaus einen überwachten Ansatz, indem wir Nachrichtenpaare als gleich oder ungleich kennzeichnen und danach strukturelle und zeitliche Merkmale extrahieren. Das Einstufungsmodell lernte anhand dieser Merkmale, insbesondere der zeitlichen Merkmale, und konnte so angelernt werden. Unsere Ergebnisse zeigen, dass das überwachte Modell eine höhere Leistung erzielen kann und somit besser geeignet ist, die oben genannten Schwierigkeiten beim Abgleich von Nachrichten zu lösen.

The Effectiveness of the Max Entropy Classifier for Feature Selection
Die Wirksamkeit des Max-Entropy-Klassifikators für die Merkmalsauswahl

Die Merkmalsauswahl ist die Aufgabe, die Anzahl der Eingangsmerkmale für eine Klassifizierung systematisch zu reduzieren. Bei der Verarbeitung natürlicher Sprache wird die grundlegende Merkmalsauswahl in der Regel durch das Auslassen gängiger Stoppwörter erreicht.

Um die Anzahl der Eingangsmerkmale noch weiter zu reduzieren, werden bei einer zahlenbasierten Eingangsdarstellung tatsächliche Merkmalsauswahlverfahren wie Transinformation oder Chi-Quadrat verwendet. Wir schlagen einen aufgabenorientierten Ansatz zur Auswahl von Merkmalen auf Grundlage von Gewichten vor, die von einem Max-Entropy-Klassifikator gelernt wurden, der für die Klassifizierung trainiert wurde.

Die restlichen Merkmale können dann von anderen Klassifikatoren genutzt werden, um die eigentliche Klassifizierung durchzuführen. Experimente mit verschiedenen Aufgaben der natürlichen Sprachverarbeitung bestätigen, dass die gewichtsbasierte Methode mit zahlenbasierten Methoden vergleichbar ist. Die Anzahl der Eingangsmerkmale kann unter Beibehaltung der Klassifizierungsleistung erheblich reduziert werden.

Impact of Anonymization on Sentiment Analysis of Twitter Postings
Auswirkung der Anonymisierung auf die Stimmungsanalyse von Twitter-Posts

Der Prozess der strategischen Modellierung und der gesamte Bereich der Strategieplanung sind komplex und stellen die Entscheidungsträger vor große Herausforderungen. Eine Herausforderung dabei ist die Einbeziehung der Bürger in den Entscheidungsprozess. Dies kann über verschiedene Formen der E-Beteiligung erfolgen, wobei die aktive/passive Bürgergewinnung eine Möglichkeit darstellt, aktuelle Diskussionen zu Themen und Problemen, die für die Allgemeinheit relevant sind, zu steuern.

Ein besseres Verständnis der Gefühle gegenüber bestimmten Themen und das daraus resultierende Verhalten der Bürger kann öffentlichen Verwaltungen neue Einsichten bieten. Gleichzeitig ist es aber wichtiger denn je, die Privatsphäre der Bürger zu respektieren, rechtskonform zu handeln und damit das Vertrauen der Öffentlichkeit zu fördern. Während die Einführung der Anonymisierung zur Gewährleistung der Wahrung der Privatsphäre eine angemessene Lösung für die genannte Probleme darstellt, ist jedoch noch unklar, ob und inwieweit sich die Anonymisierung von Daten auf die aktuellen Datenanalysetechnologien auswirkt. Daher untersucht dieser Forschungsbeitrag die Auswirkungen der Anonymisierung auf die Stimmungs-analyse in sozialen Medien im Rahmen von Smart Governance.

Drei Anonymisierungsalgorithmen werden mit Twitter-Daten getestet und die Ergebnisse werden auf Veränderungen innerhalb der resultierenden Stimmung hin analysiert. Die Ergebnisse zeigen, dass die vorgeschlagenen Anonymisierungsansätze tatsächlich einen messbaren Einfluss auf die Stimmungsanalyse haben. Dies geht sogar so weit, dass Ergebnisse für die weitere Verwendung im Bereich der strategischen Modellierung möglicherweise problematisch sein können.

Data Analytics | Modelling

A Data-Driven Approach for Detecting Autism Spectrum Disorders
Ein datengesteuerter Ansatz zur Erkennung von Autismus-Spektrum-Störungen

Autismus-Spektrum-Störungen (ASS) sind eine Gruppe von Erkrankungen, die sich durch Beeinträchtigungen der wechselseitigen sozialen Interaktion und das Vorhandensein von eingeschränkten und sich wiederholenden Verhaltensweisen kennzeichnen. Die aktuellen Mechanismen zur Erkennung von ASS sind entweder subjektiv (umfragebasiert) oder konzentrieren sich nur auf die Reaktionen auf einen einzelnen Stimulus.

In dieser Arbeit entwickeln wir maschinelle Lernmethoden zur Vorhersage von ASS auf Grundlage von Daten aus Elektrokardiogrammen (EKG) und der elektrodermalen Aktivität (EDA), die während eines sogenannten *Sensory Challenge Protocol* (SCP) gesammelt wurden, mit dem die Reaktionen auf acht Stimuli von 25 Kindern mit ASS und 25 normal entwickelten Kindern zwischen 5 und 12 Jahren beobachtet wurden. Durch die Länge der Zeitsequenzen ist es schwierig, herkömmliche maschinelle Lernalgorithmen zur Analyse dieser Datentypen zu verwenden. Stattdessen haben wir Verarbeitungs-techniken für Merkmale entwickelt, die eine effiziente Analyse der Sequenzen ohne Effektivitätsverlust ermöglichen.

Die Ergebnisse unserer Analyse der aufgezeichneten Zeitsequenzen bestätigten unsere Hypothese, dass autistische Kinder besonders stark von bestimmten sensorischen Stimuli betroffen sind. Darüber hinaus erreichte unser gemeinsames ASS-Vorhersagemodell eine Genauigkeit von 93,33 %, d. h. 13,33 % besser als das Beste aus 8 verschiedenen Basismodellen, die wir getestet haben.

Optimal Regression Tree Models through Mixed Integer Programming
Optimale Regressionsbaummodelle durch gemischt-ganzzahlige Optimierung

Die Regressionsanalyse kann zur Vorhersage von Ausgabevariablen bei einem Satz bekannter unabhängiger Variablen eingesetzt werden. Durch Regression wird eine Funktion, die die Beziehung zwischen den Variablen erfasst, an die Daten angepasst. Regressionsbaummodelle sind in der Literatur beliebt, da sie schnell berechnet werden können und einfach zu interpretieren sind.

Das Erstellen komplexer Baumstrukturen kann jedoch zu einer Überanpassung der Lerndaten und damit zu einem schlechten Vorhersagemodell führen. Diese Arbeit stellt einen Regressionsbaumalgorithmus, der anhand der mathematischen Optimierung Daten optimal in zwei Unterbereiche gliedert – sogenannte Knoten –, sowie einen statistischen Test zur Beurteilung der Qualität der Aufteilung vor. Eine Reihe von öffentlich zugänglichen Literaturbeispielen wurde verwendet, um die Leistung der Methode mit anderen, in der Literatur verfügbaren Methoden zu vergleichen.

A Spatial Data Analysis Approach for Public Policy Simulation in Thermal Energy Transition Scenarios

Ein Ansatz zur Analyse von Geodaten für die Simulation der öffentlichen Strategie in thermischen Energiewendeszenarien

Der Beitrag erläutert einen Ansatz zur Simulation der Auswirkungen öffentlicher Strategien auf die thermischen Energiewendeszenarien in städtischen Gemeinden. Der Beitrag beschreibt die zugrundeliegenden Methoden zur Berechnung des Heizenergiebedarfs von Gebäuden und die Gründe für potenzielle Zonen für thermische Energiesysteme.

Um die Auswirkungen öffentlicher Strategien auf die Gemeinden zu simulieren, entwickelten die Autoren ein räumliches, agentenbasiertes Modell, bei dem die Gebäude die Hauptobjekte sind, die sich basierend auf einer Reihe von technischen und soziodemografischen Parametern ändern können.

Um ein räumliches, agentenbasiertes Modell mit Daten zu füllen, muss eine Reihe von Open-Source- und kommerziell verfügbaren Datensätzen räumlich analysiert und zusammengeführt werden. Die ersten Ergebnisse der räumlichen, agentenbasierten Modellierung zeigen, dass öffentliche Strategien für die thermische Energiewende entsprechend simuliert werden können.

Data Analytics | Comprehensibility

A Probabilistic Approach to Web Waterfall Charts
Wahrscheinlichkeitsansatz für webbasierte Wasserfalldiagramme

Ziel dieses Beitrags ist es, einen effizienten und zuverlässigen Modellierungsansatz für probabilistische Wasserfalldiagramme zu erarbeiten, der die Zeitabläufe von webbasierten Ressourcen veranschaulicht und sich dabei besonders auf dessen Anpassung an große Datenmengen konzentriert. Eine Umsetzung mit realen Daten wird diskutiert und anhand von Beispielen veranschaulicht. Die Methode basiert auf der nichtparametrischen Dichteschätzung und wir besprechen einige subtile Aspekte, wie verrauschte Eingaben und singuläre Daten. Wir untersuchen des Weiteren Optimierungstechniken für die numerische Integration, die im Rahmen der Modellierung entsteht.

Facilitating Public Access to Legal Information: A Conceptual Model for Developing an Agile Data-driven Decision Support System
Erleichterung des öffentlichen Zugriffs auf rechtliche Informationen: Ein konzeptionelles Modell zur Entwicklung eines flexiblen, datengesteuerten Entscheidungsunterstützungssystems

Das europäische Rechtssystem ist vielschichtig und komplex. Seit seiner Einführung wurden große Mengen an Rechtsdokumenten erstellt. Dies hat erhebliche Auswirkungen auf die europäische Gesellschaft, deren verschiedene verfassungsgebende Organe regelmäßigen Zugriff auf präzise und zeitnahe rechtliche Informationen benötigen, aber häufig mit einem grundlegenden Verständnis der Rechtssprache zu kämpfen haben. Das Projekt, auf das sich dieser Beitrag konzentriert, schlägt die Entwicklung einer Reihe von nutzerzentrierten Diensten vor, die die Bereitstellung und Visualisierung von Rechtsinformationen in Echtzeit für Bürger, Unternehmen und Verwaltungen auf Grundlage einer Plattform gewährleistet, die von semantisch kommentierten Big Open Legal Data (BOLD) unterstützt wird. Ziel dieses Forschungsbeitrags ist es, durch die Entwicklung eines konzeptionellen Modells kritisch zu untersuchen, wie die aktuelle Nutzeraktivität mit den Komponenten der vorgeschlagenen Projektplattform interagiert. Aufgrund des Model Driven Design (MDD) wird die vorgeschlagene Projektarchitektur beschrieben. Sie wird durch die Anwendung des Agent Oriented Modelling (AOM) auf Grundlage von UML-(Unified Modelling Language)-Nutzeraktivitätsdiagrammen ergänzt, um sowohl die Nutzeranforderungen der vorgeschlagenen Plattform zu entwickeln als auch die Abhängigkeiten aufzuzeigen, die zwischen den verschiedenen Komponenten des vorgeschlagenen Systems bestehen.

Do we have a Data Culture?
Gibt es eine Datenkultur?

Heutzutage ist die Einführung einer „Datenkultur“ oder der „datengesteuerte“ Betrieb für viele Führungskräfte ein wünschenswertes Ziel. Was bedeutet es jedoch, wenn ein Unternehmen behauptet, eine Datenkultur zu haben? Es gibt dafür keine klare Definition. Dieser Beitrag zielt darauf ab, das Verständnis einer Datenkultur in Unternehmen zu verbessern, indem die aktuelle Verwendung des Begriffs besprochen wird. Er zeigt, dass Datenkultur eine Art Organisationskultur ist.

Eine besondere Form der Datenkultur ist die datengesteuerte Kultur. Wir kommen zu dem Schluss, dass sich eine datengesteuerte Kultur durch die Befolgung bestimmter Werte, Verhaltensweisen und Normen kennzeichnet, die eine effektive Datenanalyse ermöglichen. Neben diesen Werten, Verhaltensweisen und Normen erläutert dieser Beitrag die professionellen Rollen, die für eine datengesteuerte Kultur notwendig sind. Wir schließen die wichtige Rolle des Dateneigners ein, der die Datenkultur durch die Datenlenkung erst zu einer datengesteuerten Kultur macht. Schließlich schlagen wir eine Definition der datengesteuerten Kultur vor, die sich auf das Streben nach einer datenbasierten Entscheidungsfindung und einem ständig verbesserten Prozess der Datenanalyse konzentriert.

Dieser Beitrag unterstützt Teams und Organisationen jeder Größe, die ihre – nicht notwendigerweise großen – Datenanalysefähigkeiten verbessern möchten, indem wir auf häufig vernachlässigte, nicht-technische Anforderungen aufmerksam machen: Datenlenkung und eine geeignete Unternehmenskultur.

Short Papers

Neural Machine Translation from Natural Language into SQL with state-of-the-art Deep Learning methods
Neuronale maschinelle Übersetzung natürlicher Sprache in SQL mit modernsten Deep-Learning-Methoden

Einen Text lesen, wichtige Aussagen erkennen, zusammenfassen, Verbindungen herstellen und andere Aufgaben, die Verständnis und Kontext erfordern, sind einfach für Menschen, aber das Trainieren eines Computers in diesen Aufgaben ist eine Herausforderung. Die jüngsten Fortschritte im Bereich Deep Learning ermöglichen es, Texte tatsächlich zu interpretieren und leistungsstarke Ergebnisse bei der natürlichen Sprachverarbeitung zu erzielen. Die Interaktion mit relationalen Datenbanken über natürliche Sprache ermöglicht es Benutzern unterschiedlichster Hintergründe, große Datenmengen auf benutzerfreundliche Weise abzufragen und zu analysieren. Dieser Beitrag fasst die wichtigsten Herausforderungen und unterschiedlichen Ansätze im Zusammenhang mit Natural Language Interfaces to Databases (NLIDB) zusammen. Ein von Google entwickeltes, hochmodernes Übersetzungsmodell – Transformer – wird zur Übersetzung natürlicher Sprachabfragen in strukturierte Abfragen verwendet, um die Interaktion zwischen Benutzern und relationalen Datenbanksystemen zu vereinfachen.

Smart recommendation system to simplify projecting for an HMI/SCADA platform
Intelligentes Empfehlungssystem zur Vereinfachung der Projektierung für eine HMI/SCADA-Plattform

Die Modellierung und Verbindung von Maschinen und Hardware von Produktionsanlagen in HMI/SCADA-Softwareplattformen gilt als zeitaufwändig und erfordert Fachkenntnisse. Ein intelligentes Empfehlungssystem könnte die Projektierung unterstützen und vereinfachen. In diesem Beitrag werden überwachte Lernmethoden erörtert, um dieses Problem zu lösen. Datenmerkmale, Herausforderungen bei der Modellierung und zwei mögliche Modellierungsansätze – 1-aus-n-Code und probabilistische Themenmodellierung – werden besprochen.

Adversarial Networks — A Technology for Image Augmentation
Gegensätzliche Netzwerke – Eine Technologie zur Datenanreicherung

Eine wichtige Anwendung der Datenanreicherung ist die Unterstützung des hochmodernen maschinellen Lernens, um fehlende Werte zu ergänzen und mehr Daten aus einem bestimmten Datensatz zu generieren. Neben Methoden wie Transformation oder Patch-Extraktion können auch gegensätzliche Netzwerke genutzt werden, um die Wahrscheinlichkeitsdichtefunktion der ursprünglichen Daten zu erlernen. Mit sogenannten Generative Adversarial Networks (GANs) können neue Daten aus Rauschen generiert werden, indem ein Generator und ein Diskriminator eingesetzt werden, die in einem Nullsummenspiel versuchen, ein Nash-Gleichgewicht zu finden. Mit diesem Generator kann dann Rauschen in Erweiterungen der ursprünglichen Daten umgewandelt werden. Dieser kurze Beitrag erläutert die Verwendung von GANs, um gefälschte Bilder von Gesichtern zu erzeugen, und enthält Tipps zur Verbesserung des immer noch schwierigen Trainings von GANs.

Using supervised learning to predict the reliability of a welding process
Der Einsatz von überwachtem Lernen zur Vorhersage der Zuverlässigkeit eines Schweißprozesses

In diesem Beitrag wird überwachtes Lernen zur Vorhersage der Zuverlässigkeit von Herstellungsprozessen im industriellen Umfeld verwendet. Zur Illustration wurden die Lebensdauerdaten einer speziellen Vorrichtung aus Blech gesammelt. Es ist bekannt, dass das Schweißen der entscheidende Schritt in der Produktion ist. Um die Qualität der Schweißfläche zu prüfen, wurden mit jedem Gerät End-of-Life-Tests durchgeführt. Zur statistischen Auswertung stehen nicht nur die erfasste Lebensdauer, sondern auch Daten zur Verfügung, die das Gerät vor und nach dem Schweißprozess beschreiben, sowie Messkurven während des Schweißens, z. B. Verlauf über die Zeit. In der Regel werden die Weibull- und Log-Normalverteilung zur Modellierung der Lebensdauer verwendet. Auch in unserem Fall gelten beide als mögliche Verteilungen.

Obwohl beide Verteilungen für die Daten geeignet sind, wird die Log-Normalverteilung verwendet, da der KS-Test und der Bayes'sche Faktor etwas bessere Ergebnisse zeigen. Zur Modellierung der Lebensdauer in Abhängigkeit der Schweißparameter wird ein multivariables, lineares Regressionsmodell verwendet. Um die signifikanten Kovariablen zu finden, wird eine Mischung aus Vorwärtsauswahl und Rückwärtselimination verwendet. Mit dem T-Test wird die Wichtigkeit jeder Kovariable bestimmt, während der angepasste Determinationskoeffizient als globales Anpassungskriterium verwendet wird.

Nachdem das Modell, das die beste Anpassung bietet, bestimmt wurde, wird die Vorhersagekraft mit einer eingeschränkten Kreuzvalidierung und der Residuenquadratsumme bewertet. Die Ergebnisse zeigen, dass die Lebensdauer anhand der Schweißeinstellungen vorhergesagt werden kann. Für die Lebensdauerprognose liefert das Modell genaue Ergebnisse, wenn die Interpolation verwendet wird. Eine Extrapolation über den Bereich der verfügbaren Daten hinaus zeigt jedoch die Grenzen eines rein datengesteuerten Modells auf.

Double Blind Reviewed Full Papers

Data Analytics | Complexity

Exploring Insider Trading Within Hypernetworks

Jad Rayes
George Mason University
Fairfax, Virginia, USA
jrayes@gmu.edu

Priya Mani
George Mason University
Fairfax, Virginia, USA
pmani@masonlive.gmu.edu

***Abstract*—Insider trading can have crippling effects on the economy and its prevention is critical to the security and stability of global markets. It is hypothesized that insiders who trade at similar times share information. We analyze 400 companies and 2,000 insiders, identifying interesting trading patterns in these networks that are suggestive of illegal activity. Insiders are classified as either routine or opportunistic traders, allowing us to concentrate on well timed and highly profitable trades of the latter. Using trade classification and analyzing each trader's role in a hypernetwork, reveals cliques of opportunistic and routine traders. This idea forms the basis of a graph based detection algorithm that seeks to identify traders belonging to opportunistic cliques. The ideas of trade classification and trading cliques present interesting opportunities to develop more robust policing systems which can automatically flag illegal activity in markets, and predict the likelihood that such activity will occur in the future.**

***Index Terms*—Insider trading, anomaly detection, graph mining, hypergraphs**

I. Introduction

Insider trading is formally defined by the Securities and Exchange Commission (SEC) as the buying or selling of a security, on the basis of material, non-public information about the security. Insiders are typically company officers, board members or large shareholders within a company. In this paper we analyze the trading activity of those who engage in such insider trading.

A well known example of insider trading is the infamous case of Martha Stewart and Samuel Waksal, C.E.O of the pharmaceutical company IMClone. When it was rumored that an important drug produced by the company would not be passing FDA trials, Waksal and Stewart both sold a large portion of their shares in IMClone. Waksal dumped 5.5 million dollars worth of his shares just days before the announcement. Acting on Waksal's information, Stewart dumped a substantial portion of her shares and made over 250,000 dollars on the sale. When insiders act illegally they create an unsafe environment for other investors and undermine the public's confidence in the stock market. This has damaging affects on government institutions and businesses that rely on the stock market as a means of growth and sustenance.

Detecting illegal insider trading is not a simple task. In 2003, the daily average number of insider trades of interest to the National Association of Securities Dealers (NASD) was 5.5 million, involving over 16,000 stocks [1]. In 2015, with the cutting-edge software of the day, the Marketwatch team at the NASDAQ exchange reviewed almost 400,000 suspicious trades and had to manually refer some portion of that to the SEC for further investigation.

It is therefore imperative that we devise a method for detecting illegal insider trading on a large scale. Goldberg et al. proposed SONAR [1], a software used by the NASD that automates a large part of the detection process of suspicious insider trades. It monitors a number of stock markets and correlates daily pricing and volume data for a given stock to company-related news in order to assign likelihood scores to trades. A score reflects the probability that a trade was made illegally and if that score is above a threshold, that trade is flagged so that analysts can collect more information about it. However, SONAR appears to process all the trades of every insider. It is unable to intelligently distinguish between informative and non-informative trades, until the trade goes through its entire pipeline. Kulkarni et. al [2] first proposed the use of hypernetworks to model insider trading data. This work expands on this idea by exploring the relationships between cliques of insiders as a means of automated detection. As far as we are aware, this paper is the first to explore the significance of trading cliques and their connection to company-related news.

A. Contributions

In light of the shortcomings of existing models for detecting insider trading, we make the following contributions.

- We show that insider trading in many cases is not an individual act but involves cliques of traders.
- We show that the trades of insiders who belong to cliques are related to forthcoming company news.
- We argue that the use of hypergraphs over graphs better isolates suspicious trading patterns.
- In detecting illegal trades, we argue for the use of trade classification in order to produce more accurate results.

II. Related Work

Graph-based outlier detection is a context-sensitive problem. We might define a general graph-based anomaly detection problem as the task of identifying the graph objects (nodes, edges or substructures) that are less frequently observed than

This work was funded by the National Science Foundation. We would also like to acknowledge Dr. Carlotta Domeniconi for her guidance and help with the project.

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_1

standard objects. Due to the context-sensitive nature of outlier detection, several distinct methods have been developed. Structural methods and community methods are two such examples. Structural approaches aim to identify rare structures, in terms of either connectivity or attributes. Community-based methods try to identify members of a community who are the most different, either by connectivity or attributes [3]. These methods cannot be applied to insider trading without first developing an understanding of how the anomalies are distributed. This will heavily factor into the methodology we pursue for automated detection. So in order to understand how to approach the problem of anomaly detection, we must develop an understanding of insider trading patterns.

Insider trading has been explored in a variety of contexts. Cohen et. al perform a large scale study of insider trading, profiling a large number of insiders by their trades. They find that all insiders fall into two classifications, routine and opportunistic traders. The study concludes that opportunistic trades tend to have higher predictive power. Specifically, there is a causal relationship between opportunistic trades, market trends, and SEC prosecution activity [4]. Tamersoy et al. explore insider trading from a network centric perspective. They focus on using algorithms to construct and mine networks of traders, in order to better understand trading patterns [5].The insights gained from both studies are presented and built upon in this work. Following the work of Cohen, we want to be able to perform similar analysis on insiders in our network. Also building on the work of Tamersoy, we wish to know if there is a way to profile our traders into cliques or smaller groups where trading patterns are more routine or opportunistic respectively. This would allow us to formalize what distribution of traders are worth investigating.

We note that in several cases brought before the SEC, several insiders were prosecuted jointly, all having acted on the same information illegally. We are particularly interested in exploring the relationship between activities of trading cliques (groups of two or more traders) and company news in order to provide insight into how insiders trade.

III. Dataset

We use the same dataset in [2] but we are specifically interested in networks that are modeled as hypergraphs.

Formally, a hypergraph is a generalization of a graph $H = (X, E)$, where X is the set of vertices $\{x_1..x_{|X|}\}$ and E is a nonempty set of edges s.t. $E \subset P(X)$; $P(X)$ is the power set of X Hypergraphs capture the notion of trading cliques more accurately than traditional graphs, allowing us to focus on a pattern of trades that may reveal insight into company related news. Consider the example illustrated in Figure 1. Nodes in the network represent insiders and they are linked based on the similarity of their trading dates [2].

In the first scenario, depicted on the left, three insiders share three different trade sequences. In the second scenario, shown on the right, three insiders share the same trade sequence. If we were to model this using a graph, in both cases, we would see a single clique with three nodes. Modelling trading patterns

Fig. 1. A graph cannot distinguish between a clique and three cliques of size two

using a hypergraph would allow us to distinguish between the two different scenarios. In a hypergraph, the first scenario would be represented as three different cliques of size two and the second would be a single clique of size three. The ability to distinguish between the two different scenarios allows us to observe the significance of a sequence of trades shared by a group of traders, as opposed to just two. Importantly, paying attention to trades that follow the same chronological sequence allows us to distinguish between different types of trading behaviors.

We partition the data by company because the vast majority of insider trades tend to occur among insiders in a single company [5]. Since the reasons that insiders buy or sell stock vary, it is not helpful to represent both types of transactions in a single network. Instead, each company is partitioned into two networks. A purchase network models all of the purchases associated with that company's stock, and a sale network models all of the sales associated with that company's stock.

A hypernetwork is constructed from the data, for purchase and sale networks, for each company, as follows: Given a specific threshold t, let any number of insiders that share at least t trade dates form an edge [2]. The parameter t represents the minimum length of the longest common subsequence (LCS) of trade dates between any group of traders. Experimenting with different values of t will produce hypergraphs with different sizes for each company. Figure 2 depict the change in distribution of edges in each company's hypernetwork as we vary the threshold parameter, t. We get a larger number of graphs with over ten edges per hypernetwork when t is smaller, since more traders tend to share a smaller sequence of dates. Although choosing $t = 3$ would allow us to examine a greater number trades, significantly more cliques will involve just three trades and would make it more difficult to distinguish between routine and opportunistic trading – see IV-A. We also lose the ability to capture the rarity of larger trading cliques. Table I shows the difference between the number of all companies in the dataset, and the number of companies which have at least one clique of traders who share five trading dates. We can observe that since such trading patterns are rarer, significantly fewer companies can be modeled as hypernetworks when $t = 5$. In this paper, we adopt $t = 5$ for sale networks and $t = 10$ for purchase networks.

We also interestingly observe that sales are more common transactions (Figure 3). This is because insiders are awarded grants and equity and will often sell portions of their holdings to balance their portfolio [5].

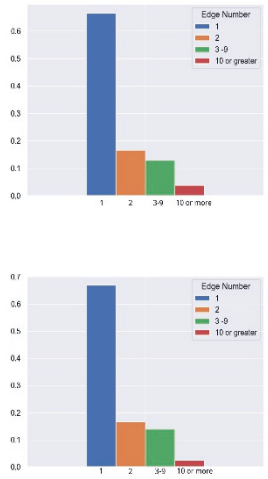


Fig. 2. A plot showing the distribution of edges across all hypergraphs in purchase networks when t = 3 (top) and t = 5 (bottom)

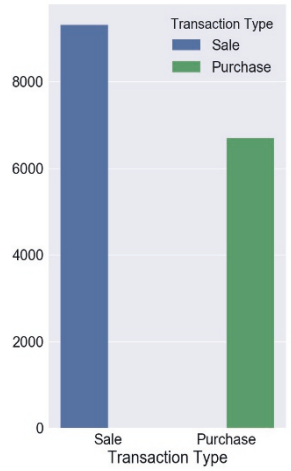


Fig. 3. A plot showing the number of purchase and sale transactions in large hypernetworks

	Companies	Insiders
GLOBAL STATISTICS	12,485	70,408
LCS STATISTICS FOR t=5	416	1,798

TABLE I
COMPARISON OF LCS-BASED HYPERNETWORK CONSTRUCTION, WITH GLOBAL DATA STATISTICS.

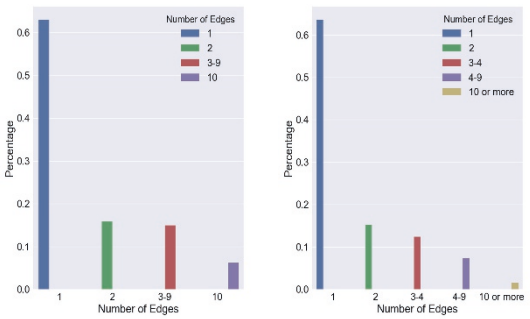


Fig. 4. The distribution of the number of hyperedges across purchase networks on the left, and sale networks on right (t = 5)

A. Preprocessing

As part of our analysis on networks, we want to consider only large networks of traders, meaning networks that comprise more than ten hyperedges. We focus on large hypernetworks, because we need enough trading cliques in order to study the connection to company-related news. Recall that we create two hypernetworks for each company, and we view each hyperedge as a trading clique, or co-occurrence of trade dates. Smaller hypergraphs are less attractive from a network-centric perspective. Figure 4 summarizes the distribution of the number of hyperedges across purchase and sale networks when $t = 5$. For example, Figure 4 depicts that majority of the hypernetworks constructed with t = 5 have just one edge, while only about 5 % of hypernetworks have ten or more edges.

Our goal is to classify all of the trades of the insiders in these larger hypernetworks and attempt to identify routine and opportunistic trading cliques, which can give valuable insight into company-related news and help explain insider trades.

IV. METHODS

A. Trade classification

Profit does not tell us the whole story. Insiders trade for a number of reasons and it is important to distinguish routine trades for diversification or liquidity from trades which are opportunely timed. These trades may not always be associated with high profit. Consider the following scenario:

Jim Jeffries, a member of the Board of Directors at ABC Corporation, buys stock in the same calendar month every year. One day, he hears rumors about an unfavorable audit of ABC Corporation. He is told that the report is about to be made public and proceeds to sell a large number of his shares the next week. The price of the stock during that week drops 3%, he incurs a 3% loss on the sale. Once the news is made public, ABC Corporation's stock plummets 12.5%. Had he sold later, he would have lost 12.5% on the sale. Jeffries clearly used his inside information about the firm to make a trade that would mitigate his losses but the sale still results in a loss.

The fact that Jim broke his usual trading patterns stands out even though he did not make a profit on the sale. Thus, it is important to look at the trading patterns of each insider in order to determine what trades are routine and what trades stand out. Several criteria can be used to classify trades as

routine or opportunistic. For example, Cohen et al. [4] classify a trader as routine if he has placed a trade in the same calendar month in the last three years. All other traders are considered opportunistic.

We adopt the methodology in [4] to classify trades as either routine or opportunistic. If the trade has occurred on the same month in at least three calendar years it is labeled as routine, otherwise it is labeled as opportunistic. The following algorithm computes the average return for opportunistic and routine trades. The signed normalized dollar amount is the profit collected from a series of transactions with the same type, price and date. It is the sum of all transactions multiplied by the stock price and normalized by the dollar volume of the stock on that date [5]. In this way, particular attention is paid to both the profit and volume of an insider trade. The algorithm presented below simply averages the profit of all opportunistic and routine trades, among all insiders in each company.

Algorithm 1 Calculate average return for routine and opportunistic trades

for each company in purchase networks, and each company in sale networks **do**

Let O denote the set of all opportunistic trades.

Let R denote the set of all routine trades.

Let $P(t_n)$ denote the profit of trade n, and it is the signed normalized dollar amount, where $-1 \leq P(t_n) \leq 1$.

Compute the average return for each trade in each respective set by

$$t_i \in O \frac{1}{|O|} \sum_{i=1}^{|O|} P(t_i)$$

$$t_j \in R \frac{1}{|R|} \sum_{j=1}^{|R|} P(t_j)$$

Table II displays a breakdown of the average profit of routine trades and opportunistic trades. Unfortunately, profit information for trades in some companies' stocks is incomplete. Consequently, we must discard one company in purchase networks and five companies in sale networks. The second column shows the average profit of all opportunistic trades as their signed normalized dollar amount. The third column shows the average profit of all routine trades, as their signed normalized dollar amount. The fourth and fifth columns are the percentage of trades which are opportunistic and routine, respectively. The last row displays the average return of opportunistic and routine trades in each company, averaged across all companies.

We observe that when all trades are considered, opportunistic purchases yield a higher return on average. We observe similar results for opportunistic sales in Table III.

Company Name	Avg. Return O	Avg. Return R	O pct.	R pct.
DNB Financial Corp.	**0.06**	0.015	28 %	72 %
Harris and Harris Group, Inc.	**0.002**	-0.025	14 %	86 %
German American Bancorp.	**.01**	-.025	5 %	95 %
NewBridge Bancorp	**-0.02**	-0.04	25 %	75 %
QCR Holdings, Inc.	**0.02**	-0.08	10 %	90 %
The York Water Company	-0.04	0.02	5 %	95 %
Consolidated Edison Inc	**0.0008**	0.0002	5 %	95 %
Southwest Georgia Financial Corporation	**0.65**	0.49	17 %	83 %
Farmers National Banc Corp	-0.03	-0.01	6 %	94 %

TABLE II
THE AVERAGE RETURN OF ROUTINE AND OPPORTUNISTIC TRADES IN PURCHASE NETWORKS

Company Name	Avg. Return O	Avg. Return R	O pct.	R pct.
Amazon, Inc.	-0.07	-0.03	9 %	91 %
Salesforce.com Inc	**0.02**	0.01	7 %	93 %
PennyMac Mortgage Investment Trust	**0.0003**	-0.02	57 %	43 %
Momenta Pharmaceuticals, Inc.	-0.05	-0.005	5 %	95 %
Workday, Inc.	**-0.08**	-0.11	40 %	60 %

TABLE III
AVERAGE RETURN OF ROUTINE AND OPPORTUNISTIC TRADES IN SALE NETWORKS

We also note that sales appear to be less routine in their timing than purchases (Figure 5). Purchases could be more routine because insiders are frequently awarded grants at the end of the fiscal year [4]. Since sales are much less predictable in their timing, we hypothesize that sales may be more strategic and hence more informative than purchases.

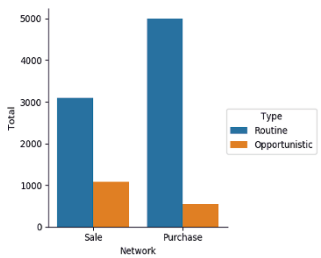


Fig. 5. Opportunistic sales are more common than Opportunistic purchases

B. Connecting trades to company-related news

We hypothesize that opportunistic trades will be more informative than routine trades. If a trade is made based on privileged information that has not become public, it is likely that after the trade is made, analyst picks and market trends will reflect the action of the insider. In order to prove this relationship between insider activity and privileged company news, we make use of the Factiva web interface. The analysis of company-related news is done manually. For each company in our dataset we identify the trades of all of the insiders. We classify each insider's trades as either routine or opportunistic using the scheme discussed in the previous subsection. For each opportunistic trade by an insider in a given company, we use the Factiva web interface to search for all news containing the company's name. We restrict the news results to be from the day of the trade to the current date. We label the results as either informative or non-informative depending on the article's pertinence and its timeliness. When reviewing an article it must meet the following criteria: The article's content must be pertinent to a subject of regulatory interest

[1]. The article should have been published in a short period following the trade, about one week. Goldberg et al. report that 85% of past insider trading cases correspond to what is known as PERM-R events. Product announcements, earnings announcements, regulatory approvals or denials, and research reports. Hence, the article's content must fall into any number of three categories for consideration:

- The company's stock is discussed in the comments of financial analysts and/or in a research report.
- The company makes headlines for reasons such as a merger, acquisition, new partnership, expansion, or new investments.
- The company issues a press release discussing either an earnings report or a change in corporate structure.

For our preliminary analysis we use companies that can be represented as hypernetworks with ten or more edges. The goal being to draw attention to routine and opportunistic cliques of traders. We define a trade to be informative if a news article published at most one week after its completion belongs to any of the three categories mentioned above. So far we have conducted analysis on all companies in purchase networks with ten or more edges, about 300 opportunistic trades spanning from 2009 to 2017. We observe that not all opportunistic trades predate pertinent company news. Figure 6 compares the percentage of opportunistic purchases and sales that are informative. We also performed preliminary analysis of routine trades on the same set of companies.

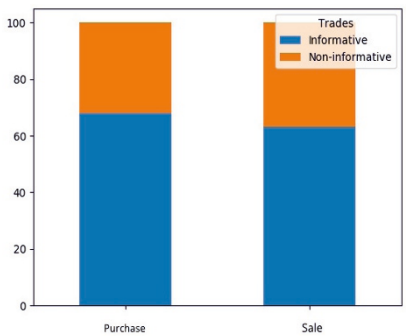


Fig. 6. Opportunistic purchases are slightly more informative than sales

Interestingly, we observed that routine trades predate company news which is generally homogeneous. For example, in DNB Financial corporation most routine trades predate a cash dividend announcement. In Harris and Harris group, most routine trades predate a quarterly letter to shareholders.

The results indicate that most opportunistic trades are highly informative and seem to reflect analyst reports and market trends that follow them. We show a breakdown of the news by category in Figure 7, which indicates that most insider trades we have observed predate company press releases.

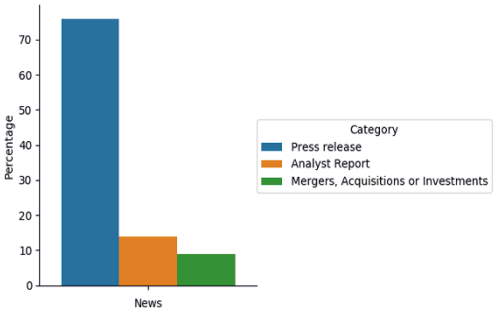


Fig. 7. A breakdown of company-related news by category

V. Results and Discussion

Using trade classification and hypergraphs to represent networks of traders we aim to isolate cliques of opportunistic traders. We are looking for a group of two or more traders who share a sequence of dates that do not fit in with their usual trading patterns. We define an opportunistic clique of traders as a group of two or more traders who share at least s opportunistic trades. For this purpose, the minimum value of $s = 1$ and s is chosen so that $s \leq t$.

These shared dates tell us two things. Firstly, the sequence of dates are likely related to company news, since a number of insiders in the same company are trading at the same time. Secondly, since the trades are opportunistic, we are operating under the assumption that the news is significant enough to pay attention to. In most cases, the trade will not coincide with a grant date or the end of the fiscal year.

Since we already constructed purchase and sale hypernetworks for each company, we aim to find smaller structures within the same networks that allow us to identify cliques of opportunistic traders. This search task involves finding a sub-graph in a hypergraph, which identifies only cliques of opportunistic traders.

Formally, if $H = (X, E)$ is the hypergraph, where X is the set of vertices and E is a nonempty set of edges s.t. $E \subset P(X)$; Then let $h = (F, Y)$ be the subgraph, where $Y \subseteq E$ and $F \subseteq X$. The goal of this algorithm is to find a subgraph h, that only contains opportunistic cliques of traders.

A. Results

We detail our results the same way as in IV-B. We analyze the dates which make up the Longest common subsequence (LCS) of each hyperedge, or clique, in the subgraph h. The intention is to see the type of company news they are placed before and how often the results are informative, according to our definition in IV-B. Since there are more opportunistic sales than opportunistic purchases, we conduct our preliminary analysis on all eight companies with large sale hypernetworks. Figure 8 shows the proportion of informative and non-informative trades in subgraphs of opportunistic cliques (a subset of all opportunistic trades) when compared against all opportunistic trades.

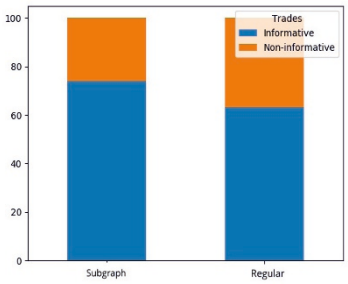


Fig. 8. A larger percentage of trades in the subgraphs are informative

Figure 9 shows the breakdown of these trades by category. Among trades in the mergers category we observe several trades which predate a merger by at least one day. We also observe a large trade on the same day as a regulatory approval – involving the company Momenta Pharmaceuticals. This approval causes that company's stock price to spike and the sale results in a big profit for the insider. Considering only opportunistic trades belonging to cliques, which is a subset of all the opportunistic trades, we observe a greater number of informative trades and several eye-catching trades which should be investigated further. We also observe from Figure 8 that significantly more trades involving opportunistic cliques are closely placed before news and analyst reports, and events such as mergers or acquisitions.

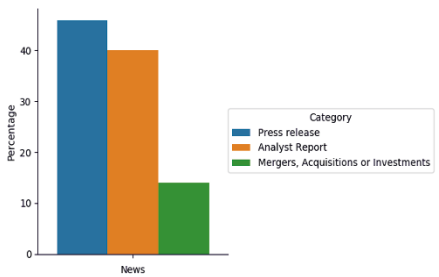


Fig. 9. Breakdown of the news by category in subgraphs

VI. Conclusions and Future Work

Domain specific considerations about insider trading is useful in the process of its detection. In fact, the application of trade classification to the problem of finding communities or structures in financial networks has proven to be a fruitful line of inquiry.

An avenue we are currently exploring is the application of network embedding techniques. Given a large network, we aim to learn feature representations that will allow us to apply clustering and classification algorithms to largely unstructured datasets with no ground truth. The idea is that, since we do not know a priori which features will be useful in the detection of insider trading, we want to produce a mathematical model that will allow us to learn them directly from the network. We are exploring the application of hypernetwork embedding algorithms like the one in [6] to this dataset. We also aim to experiment further with feature learning techniques for networks. Our current aim to is cluster learned representations for nodes in a network which can provide insight into homophily and structural roles between nodes [7]. Such representations could reveal important behavior that is critical to detecting insider trading.

References

[1] H. G. Goldberg, J. D. Kirkland, D. Lee, P. Shyr, and D. Thakker, "The nasd securities observation, new analysis and regulation system (sonar)." in *IAAI*, 2003, pp. 11–18.

[2] A. Kulkarni, P. Mani, and C. Domeniconi, "Network-based anomaly detection for insider trading," *In Proceedings of the Workshop on Inferring Networks from Non-Network Data (NETINF)*, 2017.

[3] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, vol. 29, no. 3, pp. 626–688, 2015.

[4] L. Cohen, C. Malloy, and L. Pomorski, "Decoding inside information," *The Journal of Finance*, vol. 67, no. 3, pp. 1009–1043, 2012.

[5] A. Tamersoy, B. Xie, S. L. Lenkey, B. R. Routledge, D. H. Chau, and S. B. Navathe, "Inside insider trading: Patterns & discoveries from a large scale exploratory analysis," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 797–804.

[6] A. Sharma, S. Joty, H. Kharkwal, and J. Srivastava, "Hyperedge2vec: Distributed representations for hyperedges," 2018.

[7] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

Chance influence in datasets with a large number of features

Abdel Aziz Taha
Research Studios Austria
Vienna, Austria

Alexandros Bampoulidis
Research Studios Austria
Vienna, Austria

Mihai Lupu
Research Studios Austria
Vienna, Austria

***Abstract*—Machine learning research, e.g. genomics research, is often based on sparse datasets that have very large numbers of features, but small samples sizes. Such configuration promotes the influence of chance on the learning process as well as on the evaluation. Prior research underlined the problem of generalization of models obtained based on such data. In this paper, we deeply investigate the influence of chance on classification and regression. We empirically show how considerable the influence of chance such datasets is. This brings the conclusions drawn based on them into question. We relate the observations of chance correlation to the problem of method generalization. Finally, we provide a discussion of chance correlation and guidelines that mitigate the influence of chance.**

***Index Terms*—Chance correlation, Generalization, Reproducibiliy, sparse data, Genomics**

I. Introduction

Datasets with very large numbers of features, but at the same time small numbers of samples are being frequently used in machine learning, especially in the medical domain and genomics research. A significant problem with this kind of data is the low reproducibility and generalizability of the results concluded based on it. For example, features reported by one research group as predictive regarding some outcome either widely differ from other groups' features or are not predictive when applied on other groups' data. [1] [2] [3].

Michiels et al. [4] reanalyzed seven studies that were claimed to be able to predict cancer using microarray data. They reported that the results of most of these studies are overoptimistic and five of them provide prediction methods that are not better than a random predictor. In particular, they reported instability in the feature selection in the sense that the features selected by the algorithms as predictive regarding the underlying outcomes significantly change depending on the patients considered in the training sets, such that the feature selection can be described as unrepeatable. Gene lists found by different research groups typically have only a small overlap. They also reported that this instability decreases with increasing number of samples used for training and evaluation. Jianping et al. [5] emphasized the great influence of the ratio between the number of features and the sample size on the reliability and reproducibility of prediction.

This work was carried out under support of projects VISIOMICS (FFG COIN), Safe-DEED (H2020 825225) and DMA (FFG, BMVIT 855404).

The Motivation for this research was a remarkable observation while performing experiments on an RNA microarray dataset of 16000 features (genes) of 80 diseased neuroblastoma children. The task was to predict the survival time (time until an event, e.g. relapse or death). A prediction correlation of more than 97% was achieved using a simple regression model in a cross validation experiment after performing a simple feature selection. This high accuracy was doubtful, given the small number of samples. In a follow-up experiment, we replaced all the gene data with random numbers and kept the target (survival time) unchanged. We applied exactly the same feature selection and regression algorithms. The results of the trained prediction model obtained a correlation above 95%. This remarkable observation motivated a deeper look in the influence of chance on ML models. This observation as well as the literature denoting the problem of generalizability with this kind of datasets demonstrates the need for better understanding of chance influence.

In this paper, we first demonstrate the very large influence of chance correlation on training and evaluating prediction algorithms despite using the common cross validation. We show that these results can be confirmed using thousands of random datasets with different dimensionalities and different data types for both classification and regression. We also show that the way how feature selection is performed has a great impact on chance correlation. We provide discussion of the relation between chance and the dataset dimensionality including the number of classes. Finally we conclude by providing guidelines to mitigate the influence of chance on prediction models.

II. Related Work and Background

Prior research aims at finding optimal configurations in terms of sample size as well as feature number. In this section we summarize some of this research.

Learning curves: To predict how the classification accuracy would change when the sample size is increased, it is common to use the learning curve modelling. A learning curve is a model that describes the progress of a learning process, e.g. the accuracy of a ML algorithm as a function of the number of examples fitted in the learning. A common method to implement a learning curve is to fit the inverse power law using small samples, [6], i.e.: $f(n) = an^{-\alpha} + b$, where f is

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_2

the learning rate for n samples and a, b, and α are parameters that depend on the algorithm and the dataset.

Many approaches follow this principle to predict the accuracy of an algorithm in a confidence interval around the learning curve, given a number of samples (Figueroa et al. [7]) or to estimate the minimum number of samples required for a classification to keep the error in a particular confidence interval (Mukherjee et al. [8], Dobbin et al. [9]). However, these approaches aim at optimizing the accuracy by finding the optimal number of samples and they do not consider the generalizability and the influence of chance correlation.

Bias regarding feature selection: Ambroise et al. [10] thoroughly discussed the bias caused by the feature selection method used prior to cross validation, when feature selection is performed on the entire dataset. They stated that in this case the estimation of prediction error is too optimistic, because the kind of testing is influenced by the bias stemming from the fact that the test set is a subset from set (in this case the entire set) used for feature selection. As bias correction, they suggested using a special cross-validation and bootstrapping method to correct the biased results.

Ein-Dor et al. [1] investigate the problem of robustness in the feature selection in genomics research, i.e. that genes identified as predictive regarding an outcome vary depending on the samples included in the training such that there is only a small overlap between gene lists identified using different training subsets. Their results show that thousands of samples are required to achieve an overlap of more 50% between the gene lists.

Chance Correlation: Kuligowski et al. [11] investigated the prediction accuracy in metabolomics using Partial Least Squares Discriminant Analysis. They reported that cross-validation after feature selection provides overoptimistic results due to the chance correlation. The effect of chance correlation is expressed by means of p-values calculated by using a permutation test including the variable selection, but they don't relate the chance to the number of features.

III. Notation

We provide definitions that hold for the whole paper. Two types of experiments, which will be used for analysis in this paper are defined in the following.

Definition 1: Regression based on random: Let $D = \{C_1, ..., C_m\} \cup \{C^*\}$ be a random dataset of the shape $m \times n$ where C_1 to C_m are columns (features) in the dataset and C^* is the target column. All values are of numeric data type and are generated either from a uniform or a Gaussian distribution. A regression model is trained on this random dataset to predict the target class. The predicted values are evaluated against the values of C^* to find the accuracy of the model obtained purely by chance, e.g. using the Pearson's correlation.

Definition 2: Classification from random: Let $D = \{C_1, ..., C_m\} \cup \{C^*\}$ be a random dataset of the shape $m \times n$ where C_1 to C_m are columns (features) in the dataset and C^* is the target column that partitions all n instances into r classes $Q_1 ... Q_r$. Let t_j be the true number of instances in each class Q_j. The categorical values of the features and the r classes are generated and assigned to the target randomly. A classification model is trained on the data set to predict the classes, which are then evaluated against the true classes (C^*) to find the accuracy of the classification model obtained purely by chance using some overlap metric, e.g F-Measure.

In this paper, the *shape* of a dataset is given by the number of features m, the number of samples n and the number of classes r: An $m \times n \times r$ dataset is a dataset consisting of n rows (each row referring to a data sample), m columns (each column referring to a feature) and r classes partitioning the samples into r partitions. We use the variable ρ to denote the ratio between the number of features and the number of samples, i.e. $\rho = m/n$. Furthermore, we use the term *wide dataset* to denote a dataset with $\rho > 10$.

IV. Chance influence on Prediction

In this section, we show that prediction accuracy measured for models trained using wide datasets can be for the most extent or even entirely caused by chance. We empirically quantify the extent of chance as a function of dataset shape. We also show that the way of performing feature selection as well as the validation modality, are key factors for avoiding chance influence. In particular, we generate random datasets according to Definitions 1 and 2. We train ML models (regression and classification separately), evaluate their performance and analyze the results in relation to the dataset shape and the modality of feature selection.

A. Dataset Generation

We generated three groups of random datasets, two consisting of numeric data and one consisting of categorical data. The first group (RDNUM1) contains 1000 datasets according to Definition 1, where the numbers of features as well as numbers of samples vary from 10 to 1000. The second group (RDNUM2) consists of 200 datasets according to Definition 1 with a fixed number of features, namely 10k, and sample sizes varying from 10 to 1000. This group is to represent very wide datasets like gene data. Feature values and target values in both groups are either drawn from a uniform distribution in the interval $[0, 1]$ or from a Gaussian distribution. The third group (RDCAT) consists of 1000 datasets according to Definition 2 with number of features as well as sample sizes varying from 10 to 1000. This group should represent categorical data. The feature values are either numerical like in RDNUM1 or random nominal values. The outcome is always categorical having the dummy nominal values C_0,C_1,.. C_r, where r is the number of classes, which varies from 2 to 9. All random values are generated using the random function implementation of Java 1.8.

B. Chance influence on Regression

The aim of this section is to demonstrate (i) that it is possible to train algorithms on pure random data and obtain high prediction accuracy due to chance correlation and (ii) that the way how feature selection is performed, strongly

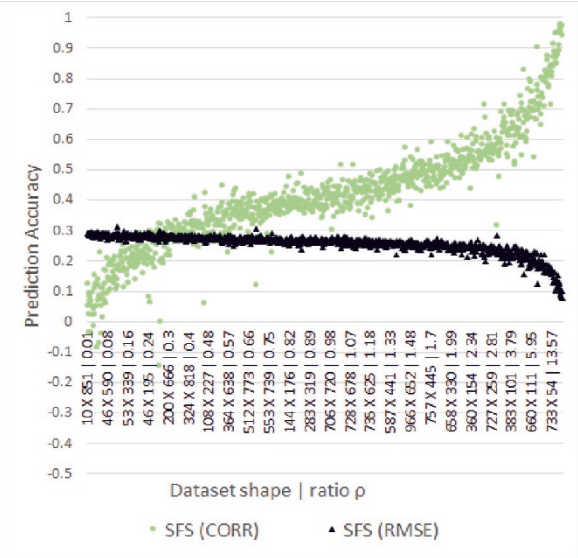


Fig. 1. Accuracy measures (Correlation and RMSE) of prediction Models trained on random datasets with different shapes using the single feature selection method (SFS). m and n vary from 10 to 1000 (Note that RMSE has here the same range $[0, 1]$ because the values in the target class are in the same interval).

affects the influence of chance on the prediction results. To this end, regression models have been trained on the datasets in the RDNUM1 group. As a feature selection, a simple BestFirst search and A Gaussian Processes classifier have been performed in a 10-fold cross validation. However, this feature selection has been performed in two different ways:

- *Single Feature Selection (SFS)*: The selection is performed only one single time prior to the cross-validation process using the whole dataset.
- *Folded Feature Selection (FFS)*: Individual feature selection for each split (fold) using the training data only (excluding the test part).

Figure 1 shows the correlation values (as accuracy measures) of 1000 prediction models, each trained on a dataset from the RDNUM1 group. It is clear that with single feature selection SFS, there is a strong increase of prediction correlation with increasing ρ (the ratio between the number of features and the number of samples). Almost perfect predictions are obtained when ρ is sufficiently high. Even when $\rho \approx 1$, there are still correlations in the order of 0.40. Figure 2 shows that the correlation with the modality Folded feature selection (FFS) are significantly lower. The vast majority of correlation values are between -0.1 and 0.1 and the RMSE values remains in the order of 0.3 which is the expectation value of RMSE for a random variable drawn from a uniform distribution in $[0, 1]$. Furthermore the accuracies have a significantly lower dependence on ρ. This is a clear indicator that the FFS feature selection modality mitigates the influence of chance and enable more reliable results.

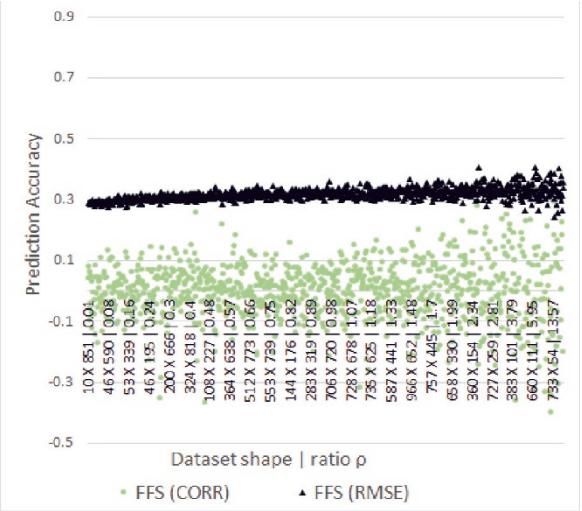


Fig. 2. Accuracy measures (Correlation and RMSE) of prediction Models trained on random datasets with different shapes using the folded feature selection method (FFS). m and n vary from 10 to 1000.

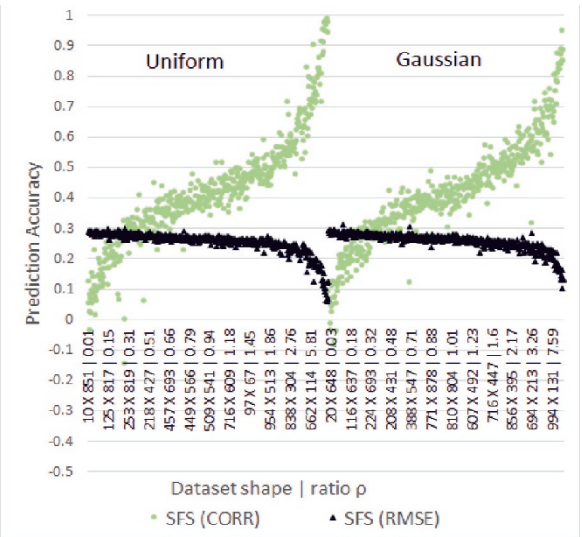


Fig. 3. Accuracy measures (Correlation and RMSE) of prediction Models trained on random datasets with different shapes using the single feature selection method (SFS) sorted according to distribution (uniform and Gaussian) and then according to ρ.

In Figure 3, the datasets are grouped by the data distribution and sorted by ρ to compare the uniform and Gaussian distributions. We see that the data distribution has almost no influence on the behavior of prediction accuracy from chance as a function of the dataset shape.

Figure 4 shows the chance behavior with very wide datasets, namely the datasets of the RDNUM2 group with 10000 features. It is remarkable to see that not only very high

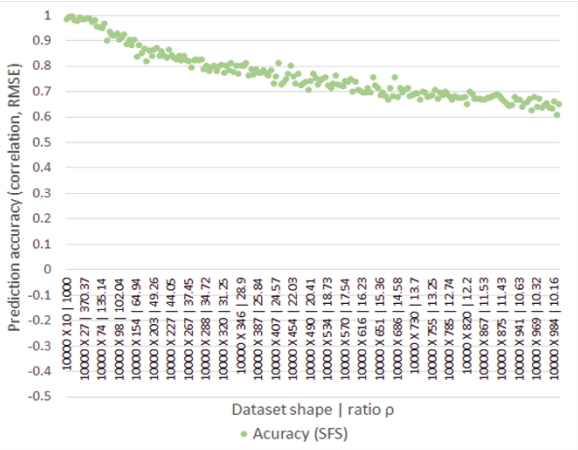


Fig. 4. Accuracy of prediction Models trained on random sets with 10k features and different sample sizes.

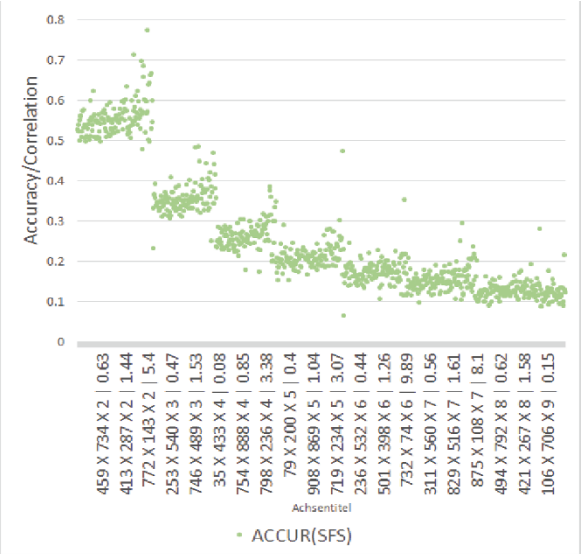


Fig. 5. Accuracy of prediction Models trained on random sets with different shapes. The datasets are first sorted according to ρ and then according to the number of classes.

accuracy can be obtained by chance with high ρ values, but that we still have accuracies around 0.5 even with sample sizes up to 1000 which significantly exceeds the typical sample size commonly used, e.g. in genomics research.

C. Chance influence on Classification

In this ssection we demonstrate the influence of chance correlation on classification, i.e. building classification model from random categorical (nominal) data. To demonstrate this, we trained J48 Trees on the 1000 datasets of RDCAT group using both SFS and FFS feature selection moralities. Figer 5 shows the accuracies of J48 trees trained trained on the datasets of the RDCAT group. The datasets are sorted at first according to ρ and then according to the number of target classes r. It is notable that the values are arranged in different levels. These levels relate to the number of target classes in the dataset.The classification accuracy strongly depend on these levels, i.e. on the number of classes r. There is already a dependence on ρ within each level, but this dependence decreases with increasing the number of classes r. It is interesting to see that there is almost no influence of ρ when there are more than 7 classes in the dataset.

V. Analysis

Observations and empirical results on random data show the considerable effect of chance on ML models. It is remarkable to see in Figure 4 that even using a sample size in the order of thousands, we still can have an additional accuracy caused by chance in the order of 60% to 70% according to state-of-the-art evaluation methods, given the number of features is such high as in the commonly used size in the genomics. In this section, we discuss the issue of chance influence from different view points in the light of the results of Section IV as well as the comments in the literature in this regards.

A. Feature Selection and chance

Feature selection methods are an essential step in ML that help exclude irrelevant features from the training process. This leads in many cases to performance enhancement, especially when the number of features is very large or when there is redundant information. Feature selection helps also as a dimensionality reduction to mitigate the effects stemming from the curse of dimensionality.

However, particularly in the case of large number of features, there are two types of features selected: (I) Features selected because they are indeed relevant, i.e. because they contain information with respect to the target class and (II) features selected because of their correlation by chance with the target class, i.e. they don't have information, but rather an accidental correlation in the underlying particular dataset. The probability of observing and selecting features of type II increases directly proportionally with the number of features and inversely proportionally with the number of instances.

Cross validation (CV) is a method to avoid over-fitting and increase generalizabilty. CV does not necessarily avoid chance influence, especially in its commonly used form, simply because the correlation stemming from chance in type II features technically does not differ from the correlation in type I features stemming from real information. However, depending on how cross validation is performed, chance influence can be considerably mitigated.

In Section IV-B, we presented two modalities of feature selection, namely the single feature selection (SFS) and the folded feature selection (FFS) and showed that the FFS considerably mitigates the chance influence, an observation that we will explain here:

- **SFS:** When a single feature selection is performed prior to a CV using the whole dataset, the feature selection step ends with a feature set that is reduced to those features that are relatively highly correlated with the target. Now splitting the dataset into folds does not change the fact that the remaining (selected) features are correlated in all instances and thus in all folds. This leads to the fact that a model trained on any nine folds will perform well on the tenth fold (assuming a 10-CV).
- **FFS:** In the folded feature selection, in each split another feature selection is performed using only the training subset. This leads to the following: Type I features selected based on the training subset will likely correlate also with the testing subset and thus lead to a higher score. On the contrary, Type II features will not correlate with the testing subset because they have accidental correlation with respect to the training subset only, thus lead to a lower score.

Of course, nothing comes without disadvantages: The result of performing a folded feature selection is different subsets of features, at least n subsets in an n-CV. This is not optimal if the aim is to identify the relevant features rather than to build a model.

B. Regression: Correlation versus RMSE

Taking a deeper look at Figure 1 and the data behind it, especially at the difference between the evaluation metrics correlation (CORR) and the root mean square error (RMSE), particularly their sensitivities to chance, one can note the following: While the CORR values span a range from 0 to 1, RMSE values remain between 0.33 and some values in the order of 0.1. The value 0.33 is the RMSE when the prediction goes totally wrong, i.e. random regarding the target class because this is the expectation value of the RMSE of uniformly distributed values in $[0, 1]$. It corresponds to zero correlation. On the oposit, the value RMSE$= 0$ corresponds to CORR$= 1$. Therefor, we normalize the RMSE values to be comparable with CORR values by $RMSE' = (0.33 - RMSE)/0.33$ to get the comparable plot in Figure 6. It shows that RMSE′ is in general less sensitive to chance than CORR: First it does not reach the extreme values (zero and one) like CORR and second RMSE′ is significantly less then CORR for the vast majority of the datasets. While CORR has a standard deviation σ^2 of 0.034, the RMSE hast a σ^2 of 0.015.

The observation above tells that the RMSE is preferable to be used as a quality measure instead of CORR, when the data is suspected to affected by chance, e.g. with a large number of features and/or a small sample size.

C. Classification: Number of classes r

Figure 5 showed the accuracies of classification models trained purely by random data. The values are clustered in different groups, where each group refers to a value of r (the number of classes), which results in different levels of accuracy. The average accuracy in each cluster (level) strongly depends on r. Actually, the minimum value in each cluster is equal to $\frac{1}{r}$, which is the probability of true assignment of an object by random, given r classes (assuming equal distribution of the objects to the classes). The increase of accuracy above this value stems from ρ, that is having additional features results in an increase of the accuracy above $\frac{1}{r}$. However, this increase is not the same in all levels, but rather decreases with increasing r, i.e. the more classes there are, the less influence ρ has on the accuracy. The influence of dimensionality ρ almost vanishes when r is more than seven. We can conclude that classification models become significantly less prone to chance influence with increasing number of classes.

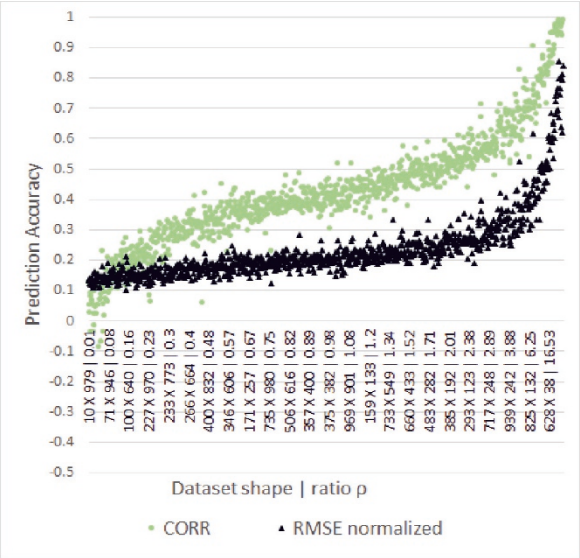


Fig. 6. RMSE normalized to be comparable with CORR. Quality measures of prediction Models trained on random datasets with different shapes using single feature selection (SFS) method. m and n vary from 10 to 1000.

D. Correction for chance

Evaluation metrics that correct for chance are not a new thing. The Cohen's Kappa metric [12] for example calculates the agreement between two raters, thereby considering the chance agreement, i.e. it corrects the agreement down based on the expected chance. The Kappa is defined as

$$Kappa = \frac{A_0 - A_e}{1 - A_e} \tag{1}$$

where A_0 is the agreement (e.g. overlap) between two rater (in our the true classification and the prediction) and A_e is the hypothetical probability of chance agreement. For a classification with n objects and r categories, A_e is defined as:

$$A_e = \frac{1}{n^2} \sum_{i=1}^{r} N_{1i} N_{2i} \tag{2}$$

where N_{ji} the number of objects predicted by rater j as Class i. For example, assume that you have n objects, where a half of them is assigned to Class 1 and the other half to Class 2. Assume that there is a dummy classifier that assigns

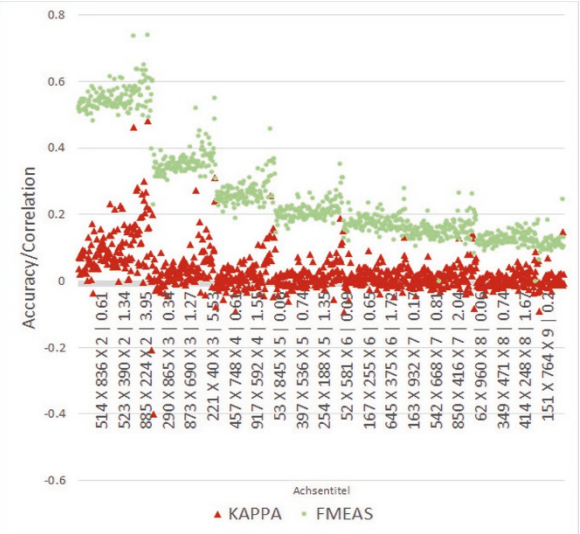


Fig. 7. Cohen's Kappa and F-measure as evaluation of prediction Models trained on the RDCAT datasets. Kappa measure corrects the values for chance by shifting them down but does not correct the accuracy increase stemming from ρ.

the classes randomly and you apply both of the F-Measure and the Kappa metrics to evaluate its accuracy. While the F-measure will score the dummy classifier with 0.5, the Kappa will score it with zero.

Equation 2 tells that Kappa considers the number of the classes as well as how the objects are assigned to the classes. It does not consider the factors that lead to these assignment. In other words, it calculates the hypothetical probability of chance just based on the object-class distribution and does not consider the number of features/instances used to find out this distribution. To demonstrate this fact, we evaluated the datasets of the RDCAT group additionally using the Kappa measure and plotted them beside the F-measure in Figure 7. If Kappa measure were able to completely correct the chance, all kappa values would be zero as expected. Kappa measure is shifted down but still has accuracy that is not corrected, namely the accuracy stemming from ρ.

This observation motivates defining a new correction for chance that additionally takes into account the number of features in relation to the number of instances under consideration of the number of classes, which is one topic of our future work.

Conclusion

We showed that in datasets with a very large number of features, like genomics datasets, chance is so considerable that it can be responsible for very high accuracies of classification and regression models. If ignored, chance could be a factor that leads to accurate, but not generalizable models. We showed that the way how feature selection is performed has a significant impact on chance influence despite cross validation, a fact that justifies to recommend using the folded feature selection within cross-validation. We also showed that the tendency of classification to be influenced by chance significantly decreases with increasing number of classes. We finally showed that different evaluation metrics are differently prone to chance. However, even the kappa metric, which is designed to correct for chance, cannot correct the chance stemming from dataset dimensionality in the sense of the ratio between number of features and sample size. These facts motivate us to continue this research to (i) formally estimate the chance in a dataset based on the dataset dimensionality expressed by the numbers of instances, features, and classes (ii) test and compare other metrics regarding their proneness to chance, (iii) extend metrics like the Kappa to consider the chance stemming from dataset dimensionality and (iv) investigate the settings that minimize the influence of chance on training and evaluation.

References

[1] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proceedings of the National Academy of Sciences*, vol. 103, no. 15, pp. 5923–5928, 2006. [Online]. Available: http://www.pnas.org/content/103/15/5923

[2] X. Fan, L. Shi, H. Fang, Y. Cheng, R. Perkins, and W. Tong, "Dna microarrays are predictive of cancer prognosis: A re-evaluation," *Clinical Cancer Research*, vol. 16, no. 2, pp. 629–636, 2010. [Online]. Available: http://clincancerres.aacrjournals.org/content/16/2/629

[3] J. P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine*, vol. 2, no. 8, 08 2005. [Online]. Available: https://doi.org/10.1371/journal.pmed.0020124

[4] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, no. 9458, pp. 488 – 492, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0140673605178660

[5] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bti171

[6] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, pp. 873–885, Aug 1989.

[7] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 8, Feb 2012. [Online]. Available: https://doi.org/10.1186/1472-6947-12-8

[8] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. R. Golub, and J. P. Mesirov, "Estimating dataset size requirements for classifying dna microarray data," *Journal of Computational Biology*, vol. 10, pp. 119–142, 2003.

[9] K. K. Dobbin, Y. Zhao, and R. M. Simon, "How large a training set is needed to develop a classifier for microarray data?" *Clinical Cancer Research*, vol. 14, no. 1, pp. 108–114, 2008. [Online]. Available: http://clincancerres.aacrjournals.org/content/14/1/108

[10] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6562–6566, 2002. [Online]. Available: http://www.pnas.org/content/99/10/6562

[11] J. Kuligowski, D. Perez-Guaita, J. Escobar, M. Guardia, M. Vento, A. Ferrer, and G. Quintas, "Evaluation of the effect of chance correlations on variable selection using partial least squares-discriminant analysis," vol. 116, pp. 835–40, 11 2013.

[12] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

Data Analytics | NLP and Semantics

Combining Lexical and Semantic Similarity Methods for News Article Matching

Mehmet Umut Sen
Sabanci University
Istanbul, Turkey

Hakki Yagiz Erdinc
Dogus University
Istanbul, Turkey

Burak Yavuzalp
Istanbul Technical University
Istanbul, Turkey

Murat Can Ganiz
Marmara University
Istanbul, Turkey

***Abstract*—Matching news articles from multiple different sources with different narratives is a crucial step towards advanced processing of online news flow. Although, there are studies about finding duplicate or near-duplicate documents in several domains, none focus on grouping news texts based on their events or sources. A particular event can be narrated from very different perspectives with different words, concepts, and sentiment due to the different political views of publishers. We develop novel news document matching method which combines several different lexical matching scores with similarity scores based on semantic representations of documents and words. Our experimental result show that this method is highly successful in news matching. We also develop a supervised approach by labeling pairs of news documents as same or not, then extracting structural and temporal features. The classification model learned using these features, especially temporal ones and train a classification model. Our results show that supervised model can achieve higher performance and thus better suited for solving above mentioned difficulties of news matching.**

***Index Terms*—news matching, machine learning, natural language processing, semantic similarity, word embeddings, jaccard similarity**

I. Introduction

The number of news from different sources and perspectives has dramatically increased due to the ever increasing variety of internet news portals and the rapid sharing of news on social media. A necessity for organizing and summarizing vast amounts of news items has emerged. Modelling news events from multiple sources would be useful for summarizing stories of long term event sequences as well as detecting false news. A necessary step of this news modelling, that we focus on in this paper, is matching news articles from different portal sources that correspond to the same event. This is an important problem since a particular event can be narrated from very different perspectives with different words, concepts, and sentiment due to the different political views of publishers in a highly polarized society.

Although, there are studies about finding duplicate or near-duplicate documents in several domains, none focus on detecting same news texts which are expressed differently. News matching can be considered as a sub-problem of semantic similarity which aims to model the similarity of different textual elements considering the meaning of those elements [1], [2]. Semantic similarity is mostly studied for the information retrieval problems such as question answering, text summarization and web search. These problems, although numerous models have been proposed and applied to them [2]–[4], are fundamentally different than the news matching problem in two ways: First, at least one of the two items in the pair are short text documents, such as questions, queries, summaries, etc. However, news articles usually are longer. Second, unlike in semantic matching problems, queries are commutative in the news matching problem, i.e. both items in the query pair are news articles.

Despite various architectures and models on different semantic similarity problems, matching long text documents is still a challenging problem [1].

We approach the problem from several aspects. First, we investigate the performance of simple lexical matching scores for the news matching problem. We argue that, even though two matched news articles have different narratives, the number of keywords and entities that define the event of concern is reasonably high for long articles and that is in favor for the lexical matching based scoring methods. We experiment on various lexical matching methods, and propose a value for threshold parameter. Following this we show an improved performance using semantic similarity which leverage word embeddings. In addition to this, combining lexical matching scores and cosine similarity scores of different word embedding methods, namely Word2Vec [5] and FastText [6]–[8], improves the performance further. These were unsupervised methods.

As mentioned before, some matched news articles may have different level of details about an event. This imbalance creates noise for the lexical matching scores. To alleviate this problem, we obtain additional scores between other fields of the two news articles, namely title and spot. We show improved performance using these additional field scores.

We also develop with supervised classification models using similarity scores as features along with other features such as the time difference and length difference of news articles. These supervised models seems especially effective.

In Section-II we summarize some previous work. In Section-III we define the problem, explain the unsupervised and supervised methods that we applied. In Section-IV we first describe the dataset that we collected, define the ex-

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_3

perimental setup, present results and discuss them. We give concluding remarks and future work at Section-VI. [1]

II. Related Work

As discussed at Section-I, most existing works on semantic similarity focus on problems such as question answering, text summarization and search [2]. Traditional methods exploit lexical databases such as WordNet [9], or any other structured semantic knowledge sources as for the biomedical text matching problem [10], [11], to measure the semantic similarity of words. However, using such external lexical resources is not practically applicable to our problem, because technical terms and named entities in the news items are of vast domain and evolve in time. In addition, high quality WordNet databases are not available in all languages.

More recent methods of semantic similarity use word embeddings that are obtained by unsupervised training on large corpora. Kenter et al. apply BM25 algorithm [12] to word embeddings, along with other meta-features, for semantic similarity of short texts [13]. We compare with this algorithm in our experiments. Kusner et al. introduce Word Mover's Distance (MVD) which casts the dissimilarity of two sets of word embeddings as Earth Mover's Distance transportation problem, where each word embedding from one document moves to embeddings in the other document with some proportions and the minimum distance is considered as the dissimilarity [14]. They report that removing one constraint in the optimization problem results in slightly lower performance but the computation time is highly reduced. We compare with this similarity algorithm, which they call Relaxed WMD (RWMD), in our experiments.

Recent supervised models mostly employ Deep Neural Network (DNN) architectures to achieve better accuracy on these tasks than the traditional methods [3], [4], [15]. Pang et al. obtain a matrix with word embeddings that contains interactions of word pairs among the two documents and treat the matrix as an image to feed into a Convolutional Neural Network (CNN) with dynamic pooling layer [3]. Hu et al. uses a similar architecture with max pooling and apply their model to various Natural Language Processing (NLP) problems such as sentence completion, matching a response to a tweet, etc. [4]. The only work, to the best of our knowledge, that deals with matching long texts of news events is the work of Liu et al. [1]. For the purpose of efficient learning with Neural Networks for long documents, they first embed these long documents into a *concept* graph, where nodes in the graph represent different concepts. Each sentence in the document is assigned to a different concept. Then, graph pairs are fed into a *Siamese Encoded Graph CNN*. They obtain better results for the news matching problem than the general similarity matching models.

These DNN models perform poorly in our problem because they require large amounts of labeled data, whereas in our dataset a small ($\approx 2K$) number of news articles are labeled. Our trials with smaller DNN models performed poorly on our dataset, therefore we discard these results from the Experiments Section.

[1]The dataset and codes are availeble at: https://github.com/donanimhaber/newsmatching

III. Method

A. Problem Definition

We tackle the problem of matching news with the same or very similar content of different portal sources. Document matches are postulated as news stories depicting the same event. We formulate the problem as inputting a pair of documents to the model and outputting a binary decision as "same" or "different".

Each document has the following fields:

1) Title: Heading of the news story.
2) Spot: Sub-heading of the story.
3) Body: Details and the main content of the story.
4) Date-time: Last modified date and time of the story.

We assume that some fields can be empty for some documents. We created another field named "text" that is the concatenation of "title", "spot" and "body" fields and used this field instead of "body". So "text" field is all the textual content of the document, and we did not use the "body" since it is usually very close to "text" because of the short lengths of "title" and "spot".

Textual fields (title_i, spot_i and text_i) of a document doc_i are sequences of words, for example $S^{(i)}_{\text{title}} = [w_0, w_1, \ldots, w_N]$.

First we describe our text based methodologies for unsupervised scoring of document pairs. In the following section, we describe the supervised setting and methods.

B. Unsupervised Scoring

Given two documents, doc_1 and doc_2, we calculate four different similarity scores for all three fields "title", "spot" and "text". Three of these scores are based on lexical matching. Another scoring uses word embeddings to capture semantic similarity.

1) Lexical Matching Scores: Jaccard Similarity Coefficient (JSC) [16] with unique words in the field is obtained as follows:

$$\text{JU}(\text{field}_i, \text{field}_j) = \frac{\left|U^{(i)}_{\text{field}} \cap U^{(j)}_{\text{field}}\right|}{\left|U^{(i)}_{\text{field}} \cup U^{(j)}_{\text{field}}\right|}, \tag{1}$$

where $U^{(i)}_{\text{field}}$ is the set of unique words in the sequence $S^{(i)}_{\text{field}}$, $|A|$ is the cardinality of a set A, $\cap$ is the intersection operator and $\cup$ is the union operator.

We calculate JSC also with all words as opposed to unique words:

$$\text{JC}(\text{field}_i, \text{field}_j) = \frac{\left|C^{(i)}_{\text{field}} \cap C^{(j)}_{\text{field}}\right|}{\left|C^{(i)}_{\text{field}} \cup C^{(j)}_{\text{field}}\right|}, \tag{2}$$

where $C^{(j)}_{\text{field}}$ is obtained by enumerating words by the occurrence count in the field. For example, for a field with words S_{field} = ["the", "cat", "sat", "on", "the", "mat"], the counted field set is S_{field} = {"the-1", "cat-1", "sat-1", "on-1", "the-2", "mat-1"} at which the second occurrence of the word "the" is now different than the first occurrence.

One issue we observed with these scorings is that for news pairs that give substantially different amount of details about an event we obtain poor JU and JC scores. Therefore we define another scoring function that replaces the denominator of the JU function with the length of the short document field:

$$\text{JS}(\text{field}_i, \text{field}_j) = \frac{\left|U^{(i)}_{\text{field}} \cap U^{(j)}_{\text{field}}\right|}{\min\left(\left|U^{(i)}_{\text{field}}\right|, \left|U^{(j)}_{\text{field}}\right|\right)}, \tag{3}$$

Even though we expect that JS would result in better similarity scores for hand matched pairs with a large difference in length, it may result in poor performance on other pairing cases, resulting in lower performance on overall dataset. But, we keep this scoring to be used as a feature with supervised classification models.

2) Word Embedding Scores: We use two different word embedding models: Word2Vec and FastText.

Word2Vec is a continuous word representation where each word is mapped to a low dimensional real vector [5]. We use the Skip-gram model which learns from large textual corpus with the objective of predicting the context words, i.e. nearby words, given an input word. A large unlabeled news articles corpus is used to train the model. After training, we obtain the vector of a given document by averaging vectors of the words contained in the document.

FastText is another continuous word representation method. Unlike previous unsupervised word embedding models such as Word2Vec [5], Doc2Vec [17], Glove [18]; FastText learns embeddings of character n-grams in addition to the word n-grams using skip-gram model [19]. It obtains the representation of a word by averaging over character n-gram vectors and the word vector. For example vector of the word "the" is obtained by averaging over the word vector "the" and the character n-gram vectors of "<t", "th", "he", "e>", "<th", "the", "he>" if only 2-grams and 3-grams are used.

Using character n-gram embeddings paves the way to dumping syntactic information into the representative vectors. This may particularly beneficial for morphologically rich languages such as Turkish, Finnish, and Hungarian. In addition, it helps to deal with syntax errors and typos which occurs frequently in non-editorial texts.

Since the number of n-grams would be enormous for a reasonably large corpus and a reasonable selection of n in the n-gram such as 3 to 6; memory requirement of this algorithm would be very high. Mikolov et al. deal with this problem by randomly grouping n-grams using a hash function and using the same vector for n-grams that are in the same group [19].

FastText obtains the document vector by first calculating word vectors, normalizing them such that their l_2 norm is 1 and then averaging them.

We use the cosine similarity to obtain the word embedding score of a pair of document fields:

$$\text{WE}(\text{field}_i, \text{field}_j) = \frac{\mathbf{v}_i^T \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}, \tag{4}$$

where $\mathbf{v}_i$ is the FastText (FT) or Word2Vec (WV) vector of field_i and $\|.\|$ is the l_2 norm.

3) Thresholding: We obtain an optimal threshold for each scoring function and for each field using a labeled dataset of news articles. We use the threshold for which precision is equal to recall since the number of negative examples, i.e. pairs that do not match, is much higher than the number of positive examples. This corresponds to the threshold for which the number of *false negatives* (falsely classified as negatives) is equal to *false positives* (falsely classified as positives).

4) Combination with Weighted Averaging: We combine different scores by weighted averaging. First, we normalize each score using the corresponding threshold and the standard deviation:

$$\text{JU}_{\text{norm}}(\text{field}_i, \text{field}_j) = \frac{\text{JU}(\text{field}_i, \text{field}_j) - \text{thr}_{\text{JU-field}}}{\text{std}(\{\text{JU}(\text{field}_i, \text{field}_j)\}_{(i,j)\in X})} \tag{5}$$

where X contains the index pairs in the training data, $\text{thr}_{\text{JU-field}}$ is the optimal threshold for the method JU and the field *field*. Normalized scores for other fields are computed similarly. After normalization of individual scores, we obtain the combined score as follows:

$$\text{COMB}(\text{field}_i, \text{field}_j) = \frac{1}{5} \sum_{m \in \{\text{JU,JS,JC,FT,WV}\}} \sum_{f \in \{\text{title,spot,text}\}} w_{\text{f}} s^{(i,j)}_{fm} \tag{6}$$

where w_f is the weight of field f independent of the scoring method, $s^{(i,j)}_{fm}$ is the normalized score of documents i and j with the method m for the field f, such as $\text{JU}_{\text{norm}}(\text{title}_i, \text{title}_j)$ for field *title* and method *JU*. We choose the weights manually proportional to the average lengths of the fields in the training data, i.e. $w_{\text{title}} < w_{\text{spot}} < w_{\text{text}}$ and weights sum to one.

Optimal threshold for the combined score is obtained similar to those of previous scoring methods. However, since the normalized scores have the optimal threshold of 0, the resulting optimal threshold for the combination is very close to 0.

5) Comparision of Semantic Similarity Methods: We compare with two similarity scoring methods which employ word embeddings. First one is the Word Mover's Distance (WMD) [14] which minimizes the following constrained optimization problem:

$$\min_{\mathbf{T}\geq 0} \sum_{i,j} T_{i,j} c(i,j) \tag{7}$$

subject to

$$\sum_{j} T_{ij} = d_i \quad \forall i \tag{8}$$

$$\sum_{i} T_{ij} = d'_j \quad \forall j \tag{9}$$

where, $c(i,j)$ is the Euclidean Distance between embeddings of the i^{th} and j^{th} unique words of the first and second document respectively, d_i and d'_j are the frequencies of the corresponding words in the document. After finding optimal $\mathbf{T}$, the distance between documents is computed by $\sum_{i,j} T_{i,j} c(i,j)$. Since solving above problem is computationally difficult, Kusner et al. relaxed this optimization problem by removing one of the constraints, obtain two easy-to-optimize problems one for each constraint (Relaxed WMD). Maximum of the two distances is used as the final distance. They show that RWMD obtains similar classification results with the WMD, therefore we obtained results using RWMD.

Kenter et al. [13] uses the following BM25 similarity algorithm [12] for short texts of S_1 and S_2:

$$f(S_1, S_2) = \sum_{w \in S_1} \text{IDF}(w) \frac{\text{sem}(w, S_2)(k_1 + 1)}{\text{sem}(w, s) + k_1(1 - b + b\frac{|S_2|}{L})} \tag{10}$$

where, $\text{IDF}(w)$ is the Inverse Document Frequency, L is the average document length, $|S_1| \geq |S_2|$, b and k_1 are meta-parameters and sem is defined as follows:

$$\text{sem}(w, S) = \max_{w' \in S} \frac{\mathbf{v}_w^T \mathbf{v}_{w'}}{||\mathbf{v}_w^T||\ ||\mathbf{x}_{w'}||} \tag{11}$$

where, $\mathbf{v}_w$ is the word embedding of the word w.

C. Supervised Classification

We use the scores described in previous section as features to be fed into a classifier. In addition, we extract *length* features and *time* features. For *length* features, we use mainly two inputs: l_1 and l_2 which represent the lengths (number of words) of fields of the document pair. We extract the following *length* features for each field ("title", "spot", "text"):

- Minimum of lengths: $\min(l_1, l_2)$
- Maximum of lengths: $\max(l_1, l_2)$
- Absolute value of difference of lengths: $|l_1 - l_2|$
- Absolute value of the difference divided by maximum of lengths $|l_1 - l_2| / \max(l_1, l_2)$
- Maximum of the lengths divided by the minimum of lengths: $\max(l_1, l_2) / \min(l_1, l_2)$

In addition to the *text length* features, we extract *time* features which are the difference of the last modified times of the two news articles. We extract two features corresponding to time difference in hours and in days. These features provide significant information to the classifier since news articles are published mostly on the same day with the event of subject.

We have 16 score features, 15 textual length features and 2 time features which amounts to a total of 33 features. These features are then fed to various classifiers including *Random Forest* (RF), *Support Vector Machines* (SVM) and *Multilayer Perceptron* (MLP). Majority Voter (MV) classifier combination results are also reported.

IV. Experiments

In subsequent sections we describe the dataset, preprocessing of the texts, parameters and details of the implementation and give results and discuss them.

A. Labeled News Dataset

We evaluate the methods on a news articles corpus in Turkish whose URLs are obtained manually by searching for the articles of the same events on different news portals. Then we crawled the web pages to obtain the fields "title", "spot", "body" and "date-time".

There are in total 20 different news portals and 2049 news items in the dataset. News articles span approximately 6 months, but majority of the articles are in a 1 month period. We obtained 693 groups of news where news in the same group correspond to the same event, i.e. positively labeled. Each group contains 2.75 documents on average. We obtained a total of 1858 positive examples and randomly choose $15,000$ negative examples. Our dataset has a total of 16858 news texts. Average numbers of words in the dataset are 6.94 ± 2.82, 25.72 ± 13.86 and 205.4 ± 223.72 for the fields *title*, *spot* and *body* respectively where second arguments are the standard deviations.

We also use an unlabeled news corpus in Turkish of size $\approx$ 4.7 million news texts for training the Word2Vec and FastText models. This unlabeled news corpus does not contain news texts of the labeled dataset. We apply the same preprocessing steps on the unlabeled as with the labeled dataset.

B. Preprocessing

We apply the following preprocessing steps to all text fields:

- Escape *html* character references.
- Remove *html* tags.
- Lowercase.
- Sentence tokenizer using NLTK toolkit [20].
- Lemmatization - Morphological analysis and disambiguation with Zemberek toolkit [21] to get lemmas.
- Stopword removel.

C. Settings

We use vector dimension of 100 for all word embeddings methods. For Word2Vec, gensim toolkit is used with skip-gram model and negative sampling [22]. Context window length parameter is chosen as 5, minimum count for filtering out words is set to 5 and training performed for 5 epochs. There were no improvement on the loss value after a few epochs.

For the FastText model; minimum count for filtering out words is set to 5, context window length parameter is chosen

as 5, bucket size is chosen to be $2,000,000$ and from 3-grams up to (including) 6-grams are used for character n-gram embeddings. Training performed for 100 epochs.

For combination of different fields and lexical matching methods as in (6), we used the following weights: $w_{\text{title}} = 0.1$, $w_{\text{spot}} = 0.3$, $w_{\text{text}} = 0.6$.

Scores for missing fields are set to the mean of the corresponding field's scores in the dataset.

We used Random Forest (RF), Multilayer Perceptron (MLP) and Support Vector Machines (SVM) for the supervised classification. Models are implemented using the *sklearn* toolkit [23]. For the RF, 100 trees are used with 2 as the minimum samples per split, 2 as the minimum samples per leaf, and Gini impurity as the split criterion. For MLP, we used 2 layers with 500 nodes each, *adam* solver, batchsize of 100 and Relu activations. Training is stopped when the loss is not decreased for at least $1e-4$ in 10 epochs. For the SVM classifier, we used *RBF* kernel and applied grid search for the penalty parameter (C) and the RBF kernel parameter (γ). We searched in $\{1e-7, 1e-5, 1e-3, 1e-1, 1, 10\}$ and $\{1, 10, 100, 1000, 10000, 100000\}$ for γ and C respectively.

We normalized all the classifier features to the range $[0, 1]$. 10-fold Cross Validation (CV) are applied to test the performances. For the feature that divides maximum length by the minimum, we used the maximum of the feature along the dataset if one of the field is missing and used 1 if both fields are missing.

For the BM25 algorithm, we applied grid search for the meta-parameters b and k_1 in $\{0.8, 1, 1.2, 1.5, 2, 2.5, 5, 10, 20, 50\}$ and $\{0, 0.001, 0.01, 0.1, 0.2, 0.5, 0.75, 0.9, 1\}$ respectively. For the RWMD method, we calculated Inverse Document Frequencies (IDF) from the large unlabeled news corpora.

V. Results

Results for lexical matching based and word embedding based similarity scoring along with compared algorithms are shown at Table-V. Here, *FT* stands for FastText, *WV* stands for Word2Vec, *COMB.* is the average of different scores as in (6).

We obtained scores for all fields *title*, *spot*, *text* along with the combination of scores for these three fields, which is depicted with *Weighted Av.* in the table, similar to using (6) except with only the related method. *Combined* results are also computed similarly.

Our results show that in general lexical matching works better than word embedding based similarity methods, except for the *title* field for which FastText works better. This shows that as the length of the text decreases, importance of semantic information increases for text matching. Another useful observations is that FastText achieves higher performance than the Word2Vec model.

Compared algorithm RWMD outperforms cosine similarity of word embeddings, but performs worse than the lexical matching based results. BM25 algorithm does not work well even though extra IDF information is incorporated.

TABLE I
F1 Results of Unsupervised Methods for Different Fields

Method	title	spot	text	Weighted Av.
RWMD-FT	0.8299	0.9386	0.9263	0.9645
RWMD-WV	0.8465	0.8762	0.9440	0.9758
BM25-FT	0.7438	0.8665	0.8881	0.9333
BM25-WV	0.7508	0.8741	0.8994	0.9413
Cosine FT	0.8407	0.9182	0.9273	0.9537
Cosine WV	0.8423	0.9225	0.9177	0.9403
JU	0.8328	**0.9535**	0.9639	0.9833
JS	0.8280	0.9476	0.9459	0.9839
JC	0.8339	0.9533	0.9709	0.9839
COMB. (FT,WM)	0.8466	0.9241	0.9225	0.9499
COMB. (JU,JS,JC)	0.8341	0.9532	**0.9817**	**0.9887**
COMB. (ALL)	**0.8617**	0.9532	0.9726	0.9833

Even though *title* and *spot* fields result in lower scores than the *text* field, score level combination of three fields (*all* in the table) achieves higher performance than the *text* field itself, even though *text* field already contains the *title* and *spot* in its content. This is expected since some news pairs have a high difference in the amount of details and titles or spots may result in noise-free match.

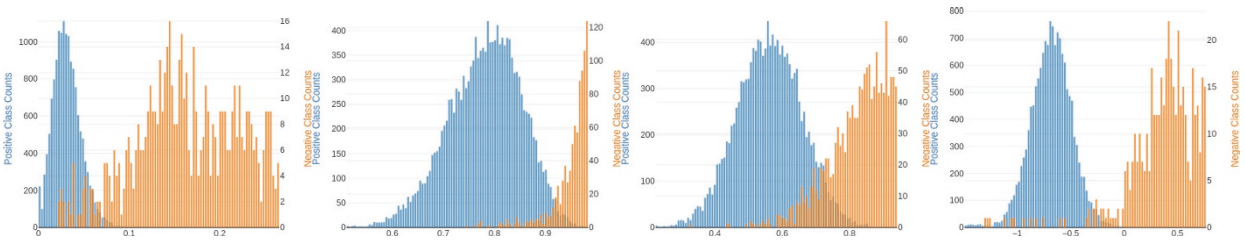
Among different lexical matching methods (*JU*, *JS*, *JC*), *JU* performs the best although results are close to each other. Results for score level combination are depicted with *COMB.* at the table. Best performing method is the score level combination of lexical matching methods, i.e. *JU*, *JS* and *JC* with the weighted averaging of fields. However, word embedding based combination works better for the *title* field.

We present histograms of scores of negative and positive pairs for some methods and fields at Figure-V. Note that we use the right y-axis for positive class for clear visualization since the number of positive examples are much lower in the dataset. We observe that lexical matching based methods result in closer to uniform histograms than word embedding based methods. However, the combined method yields more Gaussian like distribution for the positive examples. We also see higher interference between the positive and negative classes at the *title* histogram than the *text* histogram of FT.

TABLE II
Results for Supervised Methods

Method	F1	Prec.	Recall	Acc.
WE-MLP	0.9895	0.9919	0.9871	0.9977
WE-RF	0.9879	0.9876	0.9882	0.9973
WE-SVM	0.9922	0.9935	0.9909	0.9983
WE-MV	0.9922	0.9925	0.9919	0.9983
MLP	0.9925	0.9935	0.9914	0.9983
RF	0.9911	0.9914	0.9909	0.9980
SVM	0.9919	0.9935	0.9903	0.9982
MV	**0.9930**	**0.9941**	**0.9919**	**0.9985**

Results of supervised classifications are depicted at Table-V. We experimented using only word-embedding based similarity scores (FT and WV) along with additional features to test the benefit of lexical matching based scores in the supervised setting. We see improvements for all of the classifiers



(a) JU-text (b) FT-text (c) FT-title (d) CB (Combined)

Fig. 1. Score histograms of negative and positive pairs for some methods and fields

which shows the benefit of lexical matching scoring for news matching.

All supervised methods result in better performance than the best unsupervised score thresholding method (depicted as *COMB. (all)* at Table-V).

The final F1 score is 0.9930 which corresponds to falsely classifying 26 pairs in total. Some of these errors are pairs with the same narratives but for different events, such as lottery results, weather reports, etc. Another misclassification pattern is the huge difference in details between two news articles, such as a document with one sentence v.s. document with more than 100 sentences. In these cases, title matching scores are not high enough to balance the text matching scores.

VI. Conclusion

In this paper, we propose novel lexical matching based similarity calculation methods for matching long text articles of same events with different narratives. Since long articles contain higher number of event related keywords and entities than short text documents such as queries or questions, we show that, lexical matching based scores are able to obtain a fine discrimination between matched and unmatched pairings, even without the use of labels. We also proposed that obtaining scores for different fields such as title and spot of the news article would make the model more robust. Using these scores as features to be fed to classifiers, together with other length and time based features, improved the performance even further.

As a future work, we plan to work on matching two sets of news as opposed to single news, so that we can benefit from previous matchings. Another effort would be online learning of the model and testing the system in real-time.

References

[1] B. Liu, T. Zhang, D. Niu, J. Lin, K. Lai, and Y. Xu, "Matching long text documents via graph convolutional networks," *arXiv preprint arXiv:1802.07459*, 2018.

[2] Y. Fan, L. Pang, J. Hou, J. Guo, Y. Lan, and X. Cheng, "Matchzoo: A toolkit for deep text matching," *arXiv preprint arXiv:1707.07270*, 2017.

[3] L. Pang, Y. Lan, J. Guo, J. Xu, S. Wan, and X. Cheng, "Text matching as image recognition." in *AAAI*, 2016, pp. 2793–2799.

[4] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Advances in neural information processing systems*, 2014, pp. 2042–2050.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[6] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[7] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *arXiv preprint arXiv:1607.01759*, 2016.

[8] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[9] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*. Association for Computational Linguistics, 2005, pp. 13–18.

[10] T. Pedersen, S. V. Pakhomov, S. Patwardhan, and C. G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 288–299, 2007.

[11] B. T. McInnes and T. Pedersen, "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 1116–1124, 2013.

[12] S. Robertson, H. Zaragoza *et al.*, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[13] T. Kenter and M. De Rijke, "Short text similarity with word embeddings," in *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM, 2015, pp. 1411–1420.

[14] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[15] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 259–272, 2016.

[16] S. Pandit, S. Gupta *et al.*, "A comparative study on distance measuring approaches for clustering," *International Journal of Research in Computer Science*, vol. 2, no. 1, pp. 29–31, 2011.

[17] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[18] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[19] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[20] S. Bird, E. Loper, and E. Klein, "Natural language processing with python oreilly media inc," 2009.

[21] A. A. Akın and M. D. Akın, "Zemberek, an open source nlp framework for turkic languages," *Structure*, vol. 10, pp. 1–5, 2007, https://github.com/ahmetaa/zemberek-nlp.

[22] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on*

New Challenges for NLP Frameworks. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

The Effectiveness of the Max Entropy Classifier for Feature Selection

Martin Schnöll
Fact AI GmbH
Salzburg, Austria

Cornelia Ferner
Salzburg University of Applied Sciences
Salzburg, Austria

Stefan Wegenkittl
Salzburg University of Applied Sciences
Salzburg, Austria

***Abstract*—Feature selection is the task of systematically reducing the number of input features for a classification task. In natural language processing, basic feature selection is often achieved by removing common stop words. In order to more drastically reduce the number of input features, actual feature selection methods such as Mutual Information or Chi-Squared are used on a count-based input representation. We suggest a task-oriented approach to select features based on the weights as learned by a Max Entropy classifier trained on the classification task. The remaining features can then be used by other classifiers to do the actual classification. Experiments on different natural language processing tasks confirm that the weight-based method is comparable to count-based methods. The number of input features can be reduced considerably while maintaining the classification performance.**

***Index Terms*—feature selection, natural language processing, maximum entropy classification**

I. Introduction

Text classification involves two major steps: feature generation (transformation and selection) and the actual classification (training and prediction). An intuitive transformation of text into features is the Bag-of-Words (BoW) or vector space model that represents documents by word frequencies over a fixed dictionary. Frequently occurring, but purely functional terms such as articles, prepositions or conjunctions, so called stop words, are often not included in the dictionary. However, standard stop word lists[1] (e.g. [1]) have a comparably small size and little impact on the dictionary size.

A more rigorous approach to reduce the dictionary size is to apply feature selection methods that remove the lowest ranking terms based on a scoring function. Besides frequency-based ranking, scoring methods include mutual information (MI) and chi-squared ranking (χ^2) [2]. Yang et al. [3] conclude that both a k-nearest neighbor and a linear least squares classifier maintain their categorization accuracy with only 2% of the original dictionary size when features are selected based on MI and χ^2. Novakovic [4] reports the same behavior for a naive Bayes (NB) classifier, but with a less aggressive reduction of the dataset.

Classification in natural language processing tasks is commonly done by a Max Entropy (MaxEnt) classifier [5] which is trained to maximally discriminate the classes. Additionally, the learned weights per feature and class can be considered as indicator for a task-specific feature selection: Terms with absolute weights close to zero over all classes are removed from the dictionary.

We suggest using the MaxEnt classifier not only for classification, but also to systematically reduce features: Low weighting terms ("stop words") from a trained MaxEnt classifier are removed from the dictionary during preprocessing. Another classifier (MaxEnt or NB) is then trained on the reduced dictionary. Results for topic classification on document level and sentiment classification on sentence level show that the MaxEnt-based feature selection is on a par with MI and χ^2 and allows for stable classification results even with only little features left.

II. Method

An input sequence (e.g. document, sentence) is represented by a BoW vector $\vec{V} = \{v_1, \ldots, v_{|D|}\}$ defined over a fixed dictionary D. For feature selection, the MaxEnt classifier is trained to predict class $c \in C$ for a given sequence with the following probability:

$$P(c \mid \vec{V}) = \frac{\sum_{t=1}^{|D|} \exp(\lambda_{ct}\, v_t)}{\sum_{k=1}^{|C|} \sum_{t=1}^{|D|} \exp(\lambda_{kt}\, v_t)} \tag{1}$$

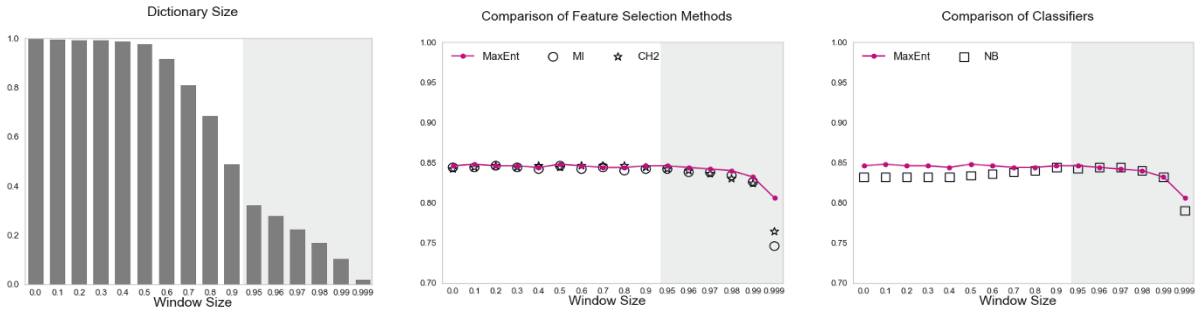
During training, the MaxEnt classifier learns to adjust the weights λ_{ct} per term t and class c. Positive weights indicate a strong correlation between term and class, while negative weights indicate an inverse correlation. A higher absolute value relates to a more significant term. Terms with absolute values close to zero across all classes thus can be considered stop words.

We suggest the following steps to systematically identify those task-specific stop words:

1) Train the MaxEnt classifier on a given task to obtain λ_{ct}.
2) Sort the term weights per class ($\lambda_{c\cdot}$) by absolute value in ascending order.
3) Removal candidates F_c^w of class c are the terms within a window of size $w \in [0, 1)$ in the sorted lists (i.e. the most uninformative $w \cdot 100\%$ of features).
4) Intersect the sets of removal candidates to obtain the list of stop words $F^w := \cap F_c^w$.

[1] For a typical list of English stop words refer to https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/feature_extraction/stop_words.py.

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_4



(a) (b) (c)

Fig. 1: Topic classification results for different window sizes w. Note the non-linear x-axis and the y-axis range. (a) Relative dictionary size after the removal of stop words. (b) Mean F1 score over 5 folds for the different feature selection methods (MI, χ^2=CH2, MaxEnt) followed by a MaxEnt classifier. (c) Mean F1 score over 5 folds for NB and MaxEnt classifier with MaxEnt-based feature selection.

The number of stop words thus depends on the window size. A term is only considered a stop word if it has a low score over all classes. Removing the stop words during preprocessing results in the new dictionary $D^w_{\text{new}} = D \setminus F^w$.

III. Experiments and Results

The effect of feature selection by removing task-specific stop words from the dataset is examined on two different tasks: topic and sentiment classification. According to [6], sentiment classification can be seen as special case of topic classification. However, it will be interesting to examine the different features that are indicative for sentiment and topic classes.

For both tasks, the data is split into two sets: one is used for computing the list of stop words, one is used for the actual classification. The classification dataset is again split into five folds for cross validation.

The experiments are implemented in Python 3.6 using the library scikit-learn [1]. Documents are vectorized by a `CountVectorizer` with a variable `stop_words` list. `MultinomialNB` is used as NB classifier, `SGDClassifier` with logarithmic `loss` as MaxEnt classifier[2]. `mutual_info_classif` and `chi2` in combination with `SelectKBest` are used for feature selection[3]. For all experiments, the mean F1 score from the `classification_report` over the folds is reported.

A. Topic Classification

The dataset used for the experiments is the 20Newsgroups corpus restricted to single labels only as provided in [7]. The task is to assign one out of 20 topics for each ”message” or document (e.g. baseball, space science or motorcycles).

A MaxEnt classifier is trained on half of the dataset to create stop word lists for different window sizes. MI and χ^2 features are also selected based on this part of the data. The dictionary size is then reduced accordingly for the second half of the dataset for classification. The original dictionary size is 70083 (with a dataset size of approx. 2.3 million tokens) and is reduced down to 1243 for $w = 0.999$. The remaining dictionary sizes corresponding to different w are shown in Figure 1a.

The mean F1 scores over five folds in Figure 1b reveal that the feature selection methods perform comparably. The MaxEnt weights are indeed a valid criterion for selecting features. Classification performance with MaxEnt-based features remains stable up to $w = 0.96$, relating to a dictionary size of 25%.

The results in Figure 1c suggest that for the topic classification task, selecting features based on MaxEnt weights is feasible, as also the NB classifier performs well with these features. While the NB classifier performs worse than the MaxEnt when trained on the full dictionary, its performance even increases for smaller dictionary sizes.

B. Sentiment Classification

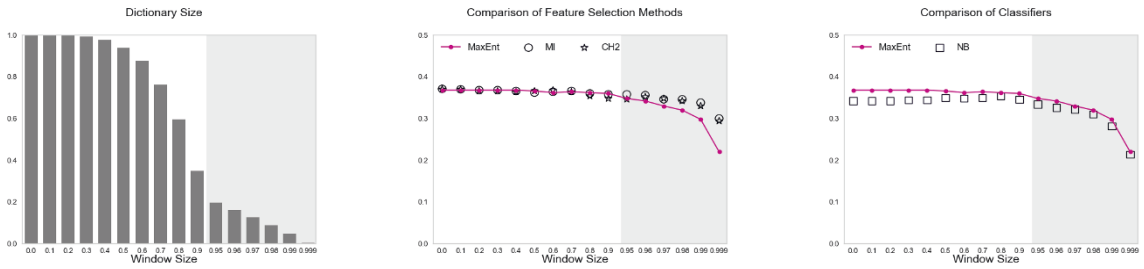
The dataset for sentiment classification is the Stanford Sentiment Treebank [8] with only full sentences used. The task is to assign one out of five sentiment classes (- -, -, ○, +, + +) per sentence. Notice the differences to the first task: less classes and much shorter input sequences. Also, the reported classification performance is lower than in the previous task, but in accordance with the original results in [8].

The original dictionary size for the sentiment classification task is 12965 (corresponding to a dataset size of 94.093 tokens) and is reduced down to 65 for the window size $w = 0.999$ (see Figure 2a).

The results are similar to the topic classification task. In Figure 2c, the performance of the NB classifier is lower for the full dictionary, but improves for smaller dictionary

[2] `coef_` stores the λ_{ct}.

[3] For a fair comparison, the number of features is determined by the remaining terms related to a given window size from the MaxEnt feature selection.



(a) (b) (c)

Fig. 2: Sentiment classification results for different window sizes w. Note the non-linear x-axis and the y-axis range. (a) Relative dictionary size after the removal of stop words. (b) Mean F1 score over 5 folds for the different feature selection methods (MI, χ^2=CH2, MaxEnt) followed by a MaxEnt classifier. (c) Mean F1 score over 5 folds for NB and MaxEnt classifier with MaxEnt-based feature selection.

space	**autos**	**hockey**	**baseball**	**religion.misc**
launch	ford	playoff	yankees	koresh
creation	bike	espn	ball	christian
sky	auto	bruins	cubs	mormons
moon	cars	nhl	phillies	promise
space	car	hockey	baseball	creation

TABLE I: Terms with highest absolute MaxEnt weights for five selected topics.

- -	-	○	+	+ +
stupid	advice	drug	decent	too
mess	human	message	chilling	masterpiece
worst	somewhere	potential	likable	refreshing
terrible	thought	inspired	fashioned	rare
dull	uninspired	white	urban	brilliant

TABLE II: Terms with highest absolute MaxEnt weights for the five sentiments.

sizes. The performance of the classifier begins to drop at a window size of $w = 0.9$ corresponding to a dictionary size of approximately 35%.

IV. Discussion and Future Work

The suggested feature selection method is applicable for both topic and sentiment classification and allows for a significant reduction of the dictionary sizes. A closer look at the top five features for each class in Tables I for topic classification and II for sentiment classification reveals that the MaxEnt classifier accurately captures important and task-specific keywords. The parts of speech differ for the two tasks: While for topic classification nouns are the predominant features, adjectives are among the top features for sentiment classification.

However, there is a tradeoff between number of features and remaining document size: Reducing the dictionary too much results in documents becoming very sparse. This is particularly true for the sentiment classification task, where the input documents are single sentences. As an example, the 65 features corresponding to a window size of $w = 0.999$ lead to 82.4% of the input sentences being empty strings.

Future experiments will investigate the suitability of the proposed feature selection method for other languages. Additionally, similar tasks in bioinformatics, such as gene selection [9], could benefit as well.

V. Conclusion

The experiments confirm that the MaxEnt classifier can as well be applied for feature selection in the context of natural language processing classification tasks. The performance is on par with other feature selection methods. Even for a significant reduction of the dictionary and corpus sizes, the performance does not suffer. However, the classification performance is affected by the remaining size of the input document.

References

[1] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[2] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval*. Cambridge University Press, 2008, vol. 39.

[3] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *International Conference on Machine Learning*, 1997.

[4] J. Novakovic, "The impact of feature selection on the accuracy of naïve bayes classifier," in *18th Telecommunications forum TELFOR*, 2010, pp. 1113–1116.

[5] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," in *Advances in Neural Information Processing Systems*, 2002, pp. 841–848.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[7] A. M. d. J. C. Cachopo, "Improving methods for single-label text categorization," *Instituto Superior Técnico, Portugal*, 2007.

[8] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 1631–1642.

[9] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, Jan 2002.

Impact of Anonymization on Sentiment Analysis of Twitter Postings

Thomas J. Lampoltshammer
Danube University Krems
Krems a. d. Donau, Austria
thomas.lampoltshammer@donau-uni.ac.at

Lőrinc Thurnay
Danube University Krems
Krems a. d. Donau, Austria
loerinc.thurnay@donau-uni.ac.at

Gregor Eibl
Danube University Krems
Krems a. d. Donau, Austria
gregor.eibl@donau-uni.ac.at

***Abstract*—The process of policy-modelling, and the overall field of policy-making are complex and put decision-makers in front of great challenges. One of them is present in form of including citizens into the decision-making process. This can be done via various forms of E-Participation, with active/passive citizen-sourcing as one way to tap into current discussions about topics and issues of relevance towards the general public. An increased understanding of feelings behind certain topics and the resulting behavior of citizens can provide great insight for public administrations. Yet at the same time, it is more important than ever to respect the privacy of the citizens, act in a legally compliant way, and therefore foster public trust. While the introduction of anonymization in order to guarantee privacy preservation represents a proper solution towards the challenges stated before, it is still unclear, if and to what extent the anonymization of data will impact current data analytics technologies. Thus, this research paper investigates the impact of anonymization on sentiment analysis of social media, in the context of smart governance. Three anonymization algorithms are tested on Twitter data and the results are analyzed regarding changes within the resulting sentiment. The results reveal that the proposed anonymization approaches indeed have a measurable impact on the sentiment analysis, up to a point, where results become potentially problematic for further use within the policy-modelling domain.**

***Index Terms*—sentiment analysis, social media, anonymization, re-identification, policy-modelling**

I. Introduction

Data have become a key asset within our society and thus it can be argued that data have even evolved to a form of social capital [1]. As a consequence, agility in regard to data analytics becomes crucial, in order to cope with current societal challenges. As these challenges effect virtually all domains, they also gained increasing importance in application scenarios within public administration, as for example, the analysis of open governmental data as pivotal element in decision-making processes [2].

Yet, the increase of data analytics also comes with an increase in complexity. This increase is also reflected in the demand of involved disciplines, in order to make the desired changes happen. While at the beginning, public administration took the required changes from a management-based perspective, this has changed to a better understanding of society, up to a point, where interdisciplinarity is not enough anymore, and the implementation of transdisciplinary approaches is more imperative than ever [3]. This demand for transdisciplinarity has led to an increasing growth of public-private partnerships, in which public administrations cooperate with private corporations and companies to tackle challenges originating from complex societal issues [4]. These partnerships can include, e.g., provision of infrastructure, consultancy, or services, such as technologies for analytical purposes. It is exactly the later field, which has been impacted by recent changes within the law of the European Union, i.e., the introduction of the General Data Protection Regulation (GDPR).

One of the biggest concerns, among others within the GDPR, is regarding privacy and the protection of personal data and information. While the term privacy is hard to define, as it strongly depends on the cultural and societal aspects within a given population context, a definition that fits the context of this research work can be found with [5, p.262], who defines privacy as "*[. . .] control over personal information*". Regarding personal data, the GDPR states the following definition [6, Article 4(1)]:

> "*personal data means any information relating to an identified or identifiable natural person (data subject); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*"

This definition implies that personal data are therefore information about an individual, which allow for identifica-

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_5

tion and thus for all other related information such as its location or means of contact [7]. From this viewpoint it can be argued, along the work in [8], that methods regarding, e.g., aggregation of personal data and re-identification, would constitute a violation of the individuals privacy. Especially, re-identification can be seen as one of the most prominent threats to individual privacy in todays data-driven society [7]. Analytic methods for big data and technologies allow for combinations of key attributes of inconspicuous and seemingly unrelated data, to re-identify individuals with ever-decreasing probabilities in regard to false positives [9]. Due to this circumstances, the GDPR foresees "*[. . .] data protection by design and by default*" [6, Article 25]. This demand is not new and has been pursued by privacy advocates for some years already, leading to the postulation of privacy preserving principles [10], [11].

However, GDPR allows for exceptions of data being protected for distinct purposes, with one exception in particular affecting public administration:

> "*processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;*" [6, Article 6(e)]

The question which arises is, if data analytics performed by the public administration or by a 3rd party contractor are not demanded by law per se, if they do fall under the mentioned exceptions of data protection and thus restrictions in processing as postulated in Article 6 of the GDPR. While an in-depth legal analysis is out of scope of this paper, the authors argue that regardless of the outcome of the before-described on-going scholar and legal discussion, public administrations are also bound in regards Article 5 (c) of GDPR , which states that [6, Article 5(c)]:

> "*[personal data shall be] adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed (data minimization)*"

The challenging part in this formulation rests within the question, how exploratory data analysis, where one does not know what data exactly to collect, to process, to analyze etc., can properly address these requirements.

A potential solution towards the before-stated problem comes in form of the anonymization of data. If such a form of data can be achieved, GDPR states [6, Article 26]:

> "*The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes.*"

Thus, the identified issues could be circumvented. Yet, by altering the data to be compliant and anonymized, another concern arises. Do the attributes of the data to be analyzed change due to the anonymization process in a way that the outcomes of the performed analysis change? Therefore, this paper is focusing exactly on this issue on the example of social media analysis, specifically on sentiment analysis of Twitter data. The associated research question thus is: To what extent does anonymization of Twitter data impact sentiment analysis performed on these data?

The remainder of this paper is structured as follows: Section 2 covers the related work, with Section 3 describing the applied methodology of paper. Section 4 afterwards presents the results of the analysis, followed by its discussion in Section 5, including existing limitations of the presented approach. The paper closes with Section 6 and the conclusions, together with an outlook for future work.

II. Related Work

A. Social Media in Public Administration

Social media are becoming more and more important in the governmental environment and the public sector in general, with an increasing importance in political participation [12] and therefore can be used as a mirror of public opinion. They provide means of new ways of interaction and communication with individuals and groups outside of the governmental space and as such, they have become an essential part in the domain of smart governance [13]. In [14], the author discusses two distinct forms of citizen-sourcing, which make explicitly use of social media, namely, active and passive citizen-sourcing. The first approach uses social media channels of governmental agencies and institutions to not only observe issues and topics of citizens, but also use these channels actively to communication with the citizens. The second approach tabs into other social media channels as well, not only the channels owned by the government and public administrations. This approach is particularly interesting, as it includes the non-directed and stimulated raw discussions of the citizens. The combination of both approaches enables access to virtually-unlimited information, directly sourced from citizens. Thereby, social media in smart governance can cover a very broad range of application scenarios [15]. In [16], the authors conducted a socio-demographic analyses of urban regions, i.e. London/England, to identify patterns and specific characteristics of Twitter users within the defined area. Such information can be useful to understand a discrete part of the population, public administration might be communicating to, or, which part of the population is reacting towards a political decision in which way. In [17] and [18], the authors investigated how citizens react towards crime news and crime events (in case of the second study regarding committed homicides) in their local area. Such information could be used by law enforcement agencies, for example, to situate awareness campaigns in areas, which are endangered by current crime waves yet are not aware of it or how to act appropriately. In [19], the authors proposed a methodology to analyses YouTube content in order to detect

online radicalization. Their approach, including topic modelling and mining of automatically-collected video metadata and manually content-based categorization of videos. The therefore gained information is not only valuable for law-enforcement agencies, but also for the respective owners of the social media platforms in order to be able to remove such videos and thus limit their spread throughout the platform as quickly as possible.

B. Sentiment Analysis in Public Administration

A particular form of content analysis, e.g. on social media, comes in form of sentiment analysis. It can be defined as: "*[. . .] the task of finding the opinions of authors about specific entities.*" [20, p.82]. The opinion of citizens is particularly interesting for policy-makers from a strategic point of view, not only during the problem definition and draft of a particular policy, but also in the later phase regarding the impact evaluation and the feedback from the citizens towards the policy as such and its actual implementation.

Several researchers have used sentiment analysis-based methodologies for their research works in the context of public administration. For example in [21], a study was conducted regarding the public mood during the Brexit voting in the United Kingdom on the basis of Twitter data. These results can help to identify, e.g., hotspots in regard to pro or contra Brexit voting. Another example can be found in the work of [22]. They analyzed, how the sentiment within U.S. local governmental tweets influences citizen involvement. The gathered insights can help to optimize governmental communication processes and policies, in order to address the citizens in a way, they feel understood and react accordingly in a positive way. A third example is given by the work of [23]. The authors also analyzed the discussions and concerns expressed by citizens on Twitter, focusing on official governmental accounts. In addition to the identification of hot topics and expressed sentiment, they used visualization techniques to reveal potential patterns within the data. These patterns can then be used to get an improved understanding of government-citizen interaction, in particular towards reactions, fear, excitement, trust etc. in the context of policy announcements.

C. Privacy Preservation and Anonymization

As already discussed in the introduction section regarding the impact of the introduction of GDPR, privacy preservation plays an important role within data analytics applications. Article 25 of GDPR therefore explicitly calls for privacy-by-design [6]. In course of this postulated paradigm, a list of eight privacy-by-design strategies haven been proposed [24].

These privacy-by-design principles and strategies point towards a specific countermeasure in regard of privacy violation, namely, anonymization. The authors in [25] mention in this context two main privacy protection models k-anonymity and -differential privacy. K-anonymity states that the chance of re-identification is below a probability of 1/k, thus, each record within the dataset has at least k-1 other records, from which it is not distinguishable [26]. Yet, in cases of records that feature sensitive attributes with a low level of variability, disclosure becomes possible [25]. Therefore, l-diverse [27] was introduced, which in theory should overcome this issue, yet also cannot prevent disclosure in case of certain types of distributions of sensitive attributes, which in turn led to the development of t-closeness [28]. This model addressed the before-mentioned issue in regard to the distribution of the sensitive attributes in such a way that it provides k-anonymity in case that the distribution of the attributes within the group in comparison to the distribution within the entire dataset is less than the defined threshold t [28].

The second privacy protection model, -differential privacy, does not even allow for the recognition of an entity being included within the data that is analyzed. Furthermore, it postulates that "*[. . .] the removal or addition of a single element in the database does not (considerably) change the results of an analysis*" and thus "*[. . .] the presence or absence of any individual on the dataset is not revealed [. . .] by the computation.*" [25, p.267]. This property triggered the initial idea of the methodology chosen by the authors of this paper, namely, to implement a naïve approach of anonymization, by altering the data to be analyzed in a way that sensitive information - that could lead to a privacy disclosure of the entity within the data is removed.

III. Methodology

The before-discussed privacy-preserving and anonymization techniques mainly focus on statistical and structured datasets, such as data in relational databases (e.g., entries per individual such as age, sex, land, income, education, etc.). Yet, not all data comes in a structured way. Especially data from social media are unstructured and often do not even follow a grammatically correct sentence structure (see Twitter data for example). Thus, the authors of this paper decided to go with an anonymization approach similar to the approach in [29], who use Named-Entity Recognition (NER) for anonymizing unstructured textual data. As it is not the focus of this paper to invent new anonymization techniques nor to improve existing ones, but to analyze the anonymization of social media data and its impact on sentiment analysis, the authors elected to implement a naïve approach towards anonymization. In the following, the methodology and data processing are described in detail.

The scenario chosen for this paper is the one of a public administration, which is currently at the starting point of the policy-cycle, namely, at the point of policy-modelling. In a first step, topics of current public discussion are identified, e.g., health-related issues, environmental changes, or tax-related issues. The public administration then collects relevant data from social media via active and/or passive citizen sourcing, e.g. Twitter.

The data for the experiment are gathered from the Internet platform Kaggle. Kaggle is a community platform for data scientists and people with an interest in data, machine learning, etc. The platform was acquired by Google/Alphabet in 2017. The platform offers, besides data, the possibility to

announce data challenges, paired with a particular dataset, and a given problem set. Individuals interested in the proposed challenges can participate with their solutions and—in case they win—receive a prize money as a reward, in exchange for granting royalty free access to the proposed solution to the issuing entity of the challenge.[1]

The entire data processing workflow can be seen in Fig. 1. In a first step (1), the authors selected a dataset [30], which contains a collection of tweets.[2]

Out of this dataset, a sub-sample of 200 tweets was chosen. The sampling was done in a random fashion. The algorithm for the random selection is based on the RAND function of Excel. Each row within the Excel represents a tweet of the entire Kaggle dataset. Via the RAND function, random Integer values are generated, as many values as there are rows within the dataset. Then, the rows are sorted in ascending order, based on the generated and line-associated Integer values. Finally, the top 200 tweets were selected to represent the sub-sample.

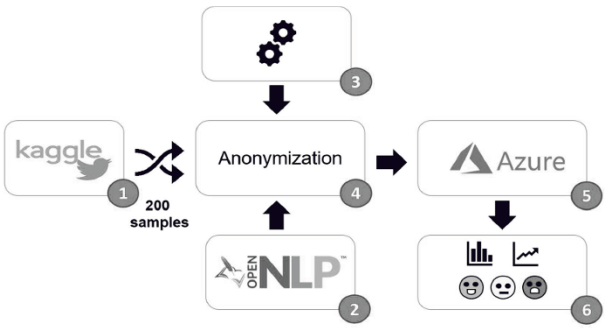


Fig. 1. Architecture of the scenario prototype.

The generated sub-sample dataset is then forwarded into the anonymization processing component (4). This software component was created by the authors and implements all data processing steps, which are required to perform the anonymization of the tweets within the sub-sample dataset. The component utilizes the Apache OpenNLP[3] library for covering all required textual data processing (2). From within this library, the Tokenizer Tools[4] are used to identify names of individuals within each tweet within the sample dataset. A tokenizer separates a sequence of characters into tokens, e.g., words, numbers etc. Therefore, the particular sequence of text, i.e. sentence(s), are checked for their boundaries, and then tokenized according to the underlying model. For the purpose of this paper, the authors chose the pre-trained model for English language, provided by the Apache OpenNLP project[5].

In total, there are three different algorithms developed by the authors with regard to the anonymization procedure (3). The first algorithm, called "Pseudo" replaces identified individuals with the fixed character sequence `Mr. Smith#` and an auto-incremented number, depending on the overall number of individuals within the text to be anonymized. For example:

"*Michael does not like Martha*" →
"*Mr. Smith#1 does not like Mr. Smith#2*"

Thus, using this approach, the structure of the sentence is not compromised, yet, the information regarding sex of the individuals is lost, respectively falsified, if one of the individuals was a female person.

The second algorithm, called "NoNNP", targets at the removal of all individuals names within the particular text. For example:

"*Michael does not like Martha*" →
"*does not like*"

In the case of this approach, the sentence structure is compromised and thus is not representing a full and syntactically correct sentence anymore.

The third and final algorithm, called "BoWwoN", uses a Bag-of-Words approach. It differs from the other two algorithms in the way, as it does not address anonymization in terms of removing the identity of individuals, but reduces the risk of re-identification, by altering the text, which in turn should make it harder to find and match towards the original data. In this approach, the sentence structure is removed completely, including all words and characters that are considered to hold no meaningful information in regard to sentiment. Such words include stop words, e.g., the, a, before, etc. as well as numbers and punctuation. As the sentence structure in this approach is also modified, this leads to a side-effect, namely, the potential loss of negations of sentiment-related words. For example:

"*Spencer is not a good guy*" →
"*Spencer good guy*"

This means that applying a Bag-of-Words approach can potentially change the sentiment within a sentence up to the point, where it changes from completely positive to completely negative and vice versa. Each of the tweets included within the sample dataset is processed within the anonymization processing component, applying all of the before-described algorithms in such a way that three new datasets emerge, each containing the results of the particular anonymization method. The results are exported into the CSV format and then forwarded to the next processing step.

The sentiment analysis component (5) is implemented with the Microsoft Azure Cognitive Services, in particular the Text Analytics tools[6]. This cloud-based analytic tool framework provides several text analysis capabilities, one of them for sentiment analysis. In a first step, the pre-processed and

[1] https://www.kaggle.com/competitions
[2] https://kaggle.com/kazanova/sentiment140
[3] https://opennlp.apache.org/
[4] https://opennlp.apache.org/docs/1.9.1/apidocs/opennlp-tools/opennlp/tools/tokenize/Tokenizer.html
[5] http://opennlp.sourceforge.net/models-1.5/
[6] https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

anonymized tweets are mapped to the required JSON format. The Azure API (Application Programming Interface) detects the language of the submitted text and performs the sentiment analysis with the according model. At the time of this work, the API supports up to 15 different languages, with English being fully-supported. The results of the sentiment analysis are again provided by the Azure API in the JSON format. These results are then re-transferred into the CSV format, for the final assessment (5) of the sentiments of all three anonymization methods, in comparison with the sentiment of the original tweets as a baseline. The sentiment values of the individual original tweets were then compared with the sentiment values of their corresponding anonymized tweets were compared to identify distortions in the sentiment values caused by the different methods of anonymization.

IV. Results

The three anonymization methods perform differently with regards to the distortion they introduce to the sentiment values of the texts they produce.

Figures 2., 3., and 4. illustrate the sentiment values of each of the 200 test cases on the y-axis. On one end of the dumbbell the original sentiment value is marked; the test cases are sorted by the original sentiment values from absolute negative 0 to absolute positive 1, forming a curve across the chart. On the other end of the dumbbell, the anonymized sentiment value of the same text case is marked. The distance between the original and anonymized sentiment values illustrate the distortion introduced by the respective anonymization method. Sentiment value 0.5 indicates neutral sentiment or that no meaningful sentiment value was possible to extract from the text.

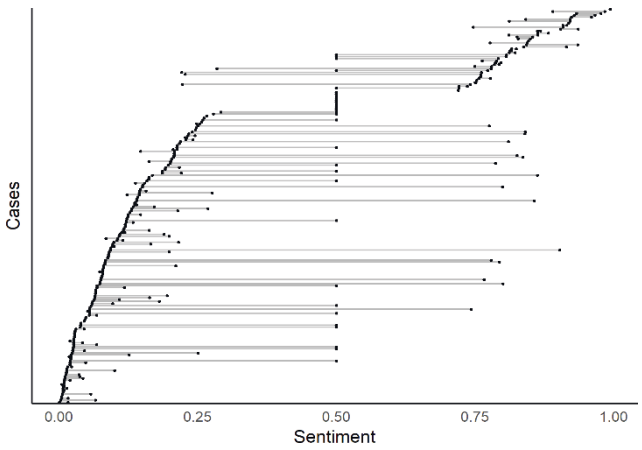


Fig. 2. Sentiment distortion of the "Pseudo" anonymization method.

All three anonymization methods introduced significant levels of sentiment distortion. A high number of long dumbbell bars indicate frequent significant distortion, whereas an increased number of 0.5 values of anonymized texts suggests a loss of sentiment value. However, to be able to

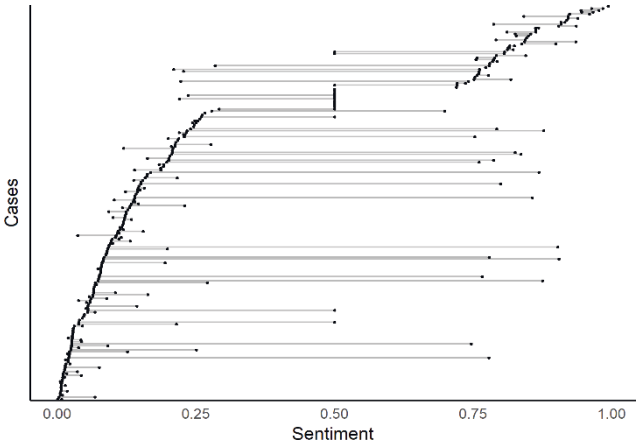


Fig. 3. Sentiment distortion of the "NoNNP" anonymization method.

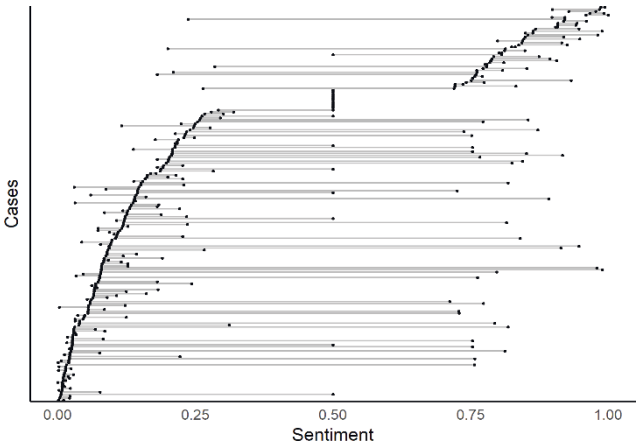


Fig. 4. Sentiment distortion of the "Bag-of-Words" anonymization method.

make an optimal choice between anonymization methods, comparable indicators must be identified. To facilitate optimal choice, the authors propose four indicators based on which the sentiment distortion properties of anonymization methods may be compared:

- $\mathbf{\Delta sent <= 0.05}$: the number of texts where sentiment distortion (i.e. the difference between the sentiment values of the original and the anonymized texts) is insignificant; not larger than 5%;
- $\mathbf{\Delta sent > 0.05}$: the number of texts where sentiment distortion is significant—larger than 5%—, however the sentiment value remains on the side of the original polarity (i.e., if the original value indicated negative sentiment ($sent_{orig} < 0.5$), the anonymized value also indicates negative sentiment, albeit with differing intensity).
- **Sentiment lost**: The number of cases where the original text has a significant sentiment value but no sentiment value can be extracted from the anonymized text

($sent_{anon} = 0.5$). The sentiment value is lost through the anonymization process. If the original text's sentiment value was also 0.5, the case is categorized as insignificant distortion.

- **Sentiment flipped**: The number of cases where the sentiment values of the original and the anonymized texts indicate opposing sentiments, regardless of their intensity, e.g. the original text communicates negative sentiment, but its anonymized versions carries positive sentiment, i.e.:

$$sent_{orig} < 0.5 \cap sent_{anon} > 0.5$$

or vice versa.

Fig. 5. illustrates how the three anonymization methods compare based on these indicators. The NoNNP method offers the most precision, it also has the lowest rate of lost sentiment value. It flips sentiment values marginally more often than the "Pseudo" method, however they perform equally in terms of significant sentiment distorsion.

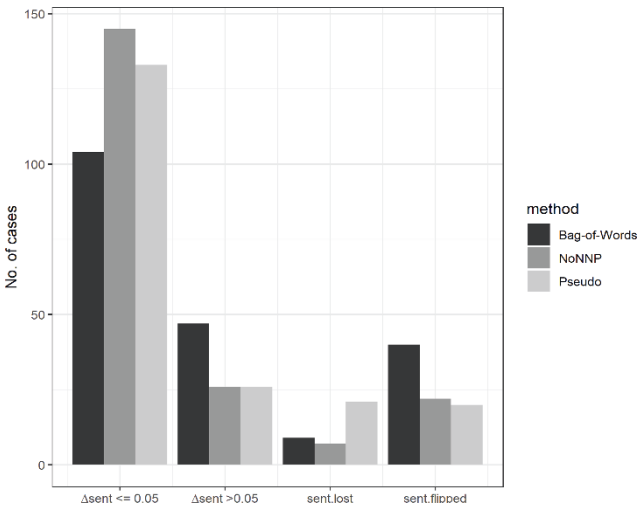


Fig. 5. Comparison of sentiment distortion properties of three anonymization methods

Fig. 6. confirms that the NoNNP method is the most precise, with the more than half of the cases having no significant deviation and shows that it fares comparably to the "Pseudo" method with regards to outliers. It also brings light to the fact that all three anonymization methods tend to err on the positive side—however this might be explained by the fact approximately three fourth of the baseline in the example texts communicate negative sentiment with different intensity, therefore any distortion is more likely to happen in the positive direction.

V. Discussion

When revisiting the research question as to what extent anonymization of Twitter data impacts sentiment analysis, the

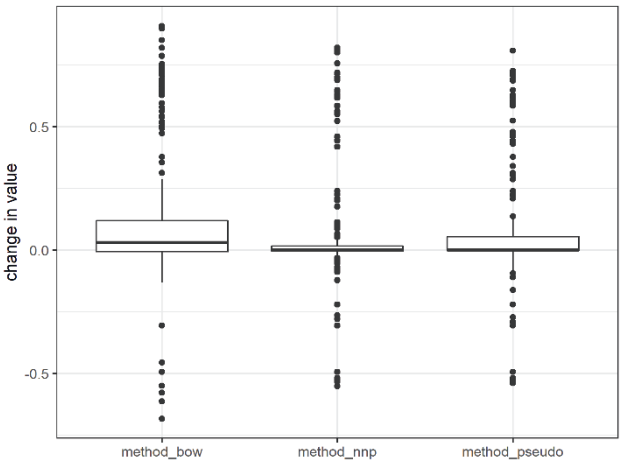


Fig. 6. Box plot of sentiment distortion of the three anonymization methods.

results have clearly demonstrated that the three applied algorithms for anonymization, "Pseudo", "NoNNP", and "BoWwoN" have a measurable impact on the sentiment analysis of the provided social media data, in this case, Twitter data. It turned out that the selected algorithms tend to shift the overall sentiment of the sample Twitter data towards a more positive picture. While the algorithms "NoNNP" and "BoWwoN" alter the sentence structure and where thus expected to show a difference, it was interesting to see that the "Pseudo" algorithm, although only replacing the identified individuals, made also an impact on the results of the sentiment analysis. The chosen implementation of the "Pseudo" algorithm uses a hashtag(#) within the replacement character sequence, yet this should not have led to any changes, as hashtags are regularly used on Twitter, also in conjunction with a reference to individuals or entities, beside the direct address via "@".

The results of this research work have to be taken with caution in regard to some associated limitations. The experiment within this paper solely focuses on Twitter data in regard to social media. Other social media data sources such as Facebook, forums, or discussion platforms might produce different results. For example, the length of the text that is going to be analyzed plays an important role in providing a better understanding of the overall context of the text. Also, there come the general issues with sentiment analysis, that is the, e.g., the detection of sarcasm [31].

Furthermore, the results are only valid in this particular setup, when referring to the Microsoft Azure platform and the included text analytics tools for the performed sentiment analysis. Other libraries might react differently, as they are based on other algorithms, which in turn can change the results of the sentiment analysis as well. Another point to consider is the selection of the three anonymization algorithms. As already discussed during the introduction of the algorithms, a naïve approach was chosen, when selecting and implementing

the three algorithms. Yet, some of the already identified issues, i.e., the loss of negations with the BoWwoN algorithm, are prominent and contribute to the positive sentiment shift. The chosen algorithms also present a simplification of the complex area of anonymization. Professional tool-kits, while often being black boxes, might produce other results as well. Finally, while the presented algorithms work well from a pure functional and technical perspective, it is still open, if this also satisfies any legal requirements for proper anonymization and thus, the establishment of sufficient means of countering re-identification risks.

VI. Conclusions

The process of policy-modelling, and the overall field of policy-making are complex issues and they pose great challenges to decision-makers. One of such challenges is including citizens into the decision-making process. This can be done via various forms of E-Participation, with active/passive citizen-sourcing as one way to tap into current discussions about topics and issues of relevance towards the general public. An increased understanding of feelings behind certain topics and the resulting behavior of citizens can provide great insight for public administrations. At the same time, it is more important than ever to respect the privacy of the citizens, act in a legally compliant way, and therefore foster public trust.

While the introduction of anonymization in order to guarantee privacy preservation represents a proper solution towards the challenges stated before, it is still unclear, if and to what extent the anonymization of data will impact current data analytic technologies. This paper analyzed the impact of anonymization on sentiment analysis of social media. It was demonstrated that already small changes towards the anonymization of data can trigger visible changes within the sentiment analysis performed via an industry standard service.

Therefore, precautions have to be taken when anonymizing data, as falsified results may have severe consequences, should policies be modelled according to them, or as the opposite, not being modelled at all, in case there was no need seen for any action by the public administration. Thus, the results of this paper should be taken further, and extensions of this work could go into several directions. First, additional anonymization algorithms could be implemented, so as to fix the existing issue with the loss of the negations in the "BoWwoN" algorithm. Furthermore, the anonymization approach could be extended by moving beyond pure detection of persons, e.g., also covering other named entities, such as locations. In addition, other means of re-identification protection could be introduced. For example, artificial "noise" could be put into the text to be analyzed. This noise could take the form of additional words that would not necessarily harm the sentence structure, but would it make more difficult to identify the original text.

Another idea, which goes into the same direction, could be the swapping of particular words with their synonyms, thereby retaining the meaning of the text. Finally, a combination of several of these algorithms could be interesting in order to find out whether the desired effect of anonymization can be strengthened or—if some approaches would even nullify themselves—reducing the wanted effect to zero.

Another aspect comes in form of other libraries or services for sentiment analysis. Microsofts Azure platform is only one available online service among others, e.g., IBM Watson. But there also exist sentiment algorithms as stand-alone libraries or as algorithms being implemented into data processing tools, such as the statistic suite R. These services and tools could be used to perform the same analytic process and afterwards compare the results in terms of robustness of each particular service/library in terms of the modifications made by the proposed anonymization algorithms.

Finally, an extensive legal analysis could complement the before-described technical extension. While the suggested solutions might work from a pure functional point of view, it is still an open question whether they would qualify in terms of GDPR compliance from a legal perspective. Thus, the legal analysis could go into details in regard to coverage of the proposed solutions of various aspects of GDPR and the privacy-by-design principles.

Acknowledgment

The research leading to these results is developed in the context of the SmartGov Project (Advanced decision support for Smart Governance). It has received funding from the Joint Programming Initiative (JPI) Urban Europe, i.e., the program ERA-NET Cofund Smart Cities and Communities (ENSCC), under the European Unions Horizon 2020 program.

References

[1] T. J. Lampoltshammer and J. Scholz, "Open Data as Social Capital in a Digital Society," in *Rethinking Social Capital: Global Contributions from Theory and Practice*, E. Kapferer, I. Gstach, A. Koch, and C. Sedmak, Eds. Newcastle upon Tyne: Cambridge Scholars Publishing, 2017, pp. 137–150.

[2] Y. Charalabidis, A. Zuiderwijk, C. Alexopoulos, M. Janssen, T. Lampoltshammer, and E. Ferro, *The World of Open Data - Concepts, Methods, Tools and Experiences*. Springer, 2018.

[3] A. Collin, "Multidisciplinary, interdisciplinary, and transdisciplinary collaboration: Implications for vocational psychology," *International Journal for Educational and Vocational Guidance*, vol. 9, no. 2, pp. 101–110, 2009.

[4] H. Wang, W. Xiong, G. Wu, and D. Zhu, "Public–private partnership in public administration discipline: a literature review," *Public Management Review*, vol. 20, no. 2, pp. 293–316, 2018.

[5] L. D. Introna, "Privacy and the computer: why we need privacy in the information society," *Metaphilosophy*, vol. 28, no. 3, pp. 259–275, 1997.

[6] European Commission, "Regulation (Eu) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/Ec (General Data Protection Regulation)," 2016.

[7] E. Politou, E. Alepis, and C. Patsakis, "Forgetting personal data and revoking consent under the gdpr: Challenges and proposed solutions," *Journal of Cybersecurity*, vol. 4, no. 1, p. tyy001, 2018.

[8] D. J. Solove, "A taxonomy of privacy," *U. Pa. L. Rev.*, vol. 154, p. 477, 2005.

[9] A. Narayanan and V. Shmatikov, "Myths and fallacies of personally identifiable information," *Communications of the ACM*, vol. 53, no. 6, pp. 24–26, 2010.

[10] A. Cavoukian, "Privacy by design: origins, meaning, and prospects for assuring privacy and trust in the information era," in *Privacy protection measures and technologies in business organizations: aspects and standards*. IGI Global, 2012, pp. 170–208.
[11] M. Langheinrich, "Privacy by designprinciples of privacy-aware ubiquitous systems," in *International conference on Ubiquitous Computing*. Springer, 2001, pp. 273–291.
[12] R. Effing, J. Van Hillegersberg, and T. Huibers, "Social media and political participation: are facebook, twitter and youtube democratizing our political systems?" in *International conference on electronic participation*. Springer, 2011, pp. 25–35.
[13] J. I. Criado, R. Sandoval-Almazan, and J. R. Gil-Garcia, "Government innovation through social media," 2013.
[14] E. N. Loukis, "Citizen-sourcing for public policy making: Theoretical foundations, methods and evaluation," in *Policy Analytics, Modelling, and Informatics*. Springer, 2018, pp. 179–203.
[15] A. L. Kavanaugh, E. A. Fox, S. D. Sheetz, S. Yang, L. T. Li, D. J. Shoemaker, A. Natsev, and L. Xie, "Social media use by government: From the routine to the critical," *Government Information Quarterly*, vol. 29, no. 4, pp. 480–491, 2012.
[16] B. Hofer, T. J. Lampoltshammer, and M. Belgiu, "Demography of twitter users in the city of london: An exploratory spatial data analysis approach," in *Modern Trends in Cartography*. Springer, 2015, pp. 199–211.
[17] T. J. Lampoltshammer, O. Kounadi, I. Sitko, and B. Hawelka, "Sensing the public's reaction to crime news using the links correspondence method," *Applied geography*, vol. 52, pp. 57–66, 2014.
[18] O. Kounadi, T. J. Lampoltshammer, E. Groff, I. Sitko, and M. Leitner, "Exploring twitter to analyze the publics reaction patterns to recently reported homicides in london," *PloS one*, vol. 10, no. 3, p. e0121848, 2015.
[19] S. Agarwal and A. Sureka, "Topic-specific youtube crawling to detect online radicalization," in *International Workshop on Databases in Networked Information Systems*. Springer, 2015, pp. 133–151.
[20] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.
[21] T. Lansdall-Welfare, F. Dzogang, and N. Cristianini, "Change-point analysis of the public mood in uk twitter during the brexit referendum," in *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 434–439.
[22] S. M. Zavattaro, P. E. French, and S. D. Mohanty, "A sentiment analysis of us local government tweets: The connection between tone and citizen involvement," *Government Information Quarterly*, vol. 32, no. 3, pp. 333–341, 2015.
[23] R. B. Hubert, E. Estevez, A. Maguitman, and T. Janowski, "Examining government-citizen interactions on twitter using visual and sentiment analysis," in *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*. ACM, 2018, p. 55.
[24] J.-H. Hoepman, *Privacy Design Strategies (The Little Blue Book)*. Nijmegen: Radboud University, 2018.
[25] J. Salas and J. Domingo-Ferrer, "Some basics on privacy techniques, anonymization and their big data challenges," *Mathematics in Computer Science*, pp. 1–12, 2018.
[26] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," technical report, SRI International, Tech. Rep., 1998.
[27] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "\ell-diversity: Privacy beyond\kappa-anonymity," in *null*. IEEE, 2006, p. 24.
[28] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007, pp. 106–115.
[29] F. Hassan, J. Domingo-Ferrer, and J. Soria-Comas, "Anonymization of unstructured data via named-entity recognition," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2018, pp. 296–305.
[30] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, vol. 1, no. 12, 2009.
[31] D. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis," in *LREC 2014 Proceedings*. ELRA, 2014.

Data Analytics | Modelling

A Data-Driven Approach for Detecting Autism Spectrum Disorders

Manika Kapoor
San José State University
San Jose, CA, USA
manika.kapoor@sjsu.edu

David C. Anastasiu*
San José State University
San Jose, CA, USA
david.anastasiu@sjsu.edu

***Abstract*—Autism spectrum disorders (ASDs) are a group of conditions characterized by impairments in reciprocal social interaction and by the presence of restricted and repetitive behaviors. Current ASD detection mechanisms are either subjective (survey-based) or focus only on responses to a single stimulus. In this work, we develop machine learning methods for predicting ASD based on electrocardiogram (ECG) and skin conductance (SC) data collected during a sensory challenge protocol (SCP) in which the reactions to eight stimuli were observed from 25 children with ASD and 25 typically developing children between 5 and 12 years of age. The length of the time series makes it difficult to utilize traditional machine learning algorithms to analyze these types of data. Instead, we developed feature processing techniques which allow efficient analysis of the series without loss of effectiveness. The results of our analysis of the protocol time series confirmed our hypothesis that autistic children are greatly affected by certain sensory stimulation. Moreover, our ensemble ASD prediction model achieved 93.33% accuracy, which is 13.33% higher than the best of 8 different baseline models we tested.**

***Index Terms*—autism spectrum disorders, large time series, data-driven autism prediction, feature extraction from time series**

I. Introduction

Autism spectrum disorders (ASDs) are conditions which can lead to impairments in reciprocal social interaction and communication, and restricted and repetitive behaviors in subjects. These neurodevelopmental disorders do not have a cure, but their early detection increases the chances of patients being able to develop coping mechanisms that improve their ability to function in society. Current ASD detection mechanisms are focused on the observation of a subject's social interaction. The instruments used for such assessments are lengthy and require extensive training, which prevents them from being used on the overall population. Before referring the subjects for further evaluation, they are first identified as at-risk via a screening process which is sometimes not accurate [1]. The social responsiveness scale (SRS) test, the most popular of such screening instruments, was shown to only have 0.78 sensitivity and 0.67 specificity [2]. Recent work has identified autonomic and behavioral responses of children with autism to be different from those of typically developing (TD) children in response to auditory [3] or visual stimuli [4].

Our research project utilizes longitudinal physiological data collected from multiple sensors in response to a protocol involving eight stimuli sequentially administered to a mixed group of ASD and TD children. Each protocol took approximately one hour to execute and resulted in large amounts of time series data consisting of millions of correlated values across the length of the protocol. A subject may be affected by one stimulus and its residual effect may be present during the administration of the next stimulus, which suggests the sensor data should be analyzed as a time-dependent series. However, analyzing such large time series is a challenging task, both in terms of the time and the space requirements of the time series analysis methods. In our research, we develop several feature extraction techniques that transform the time series into a form which can be used for efficient analysis and prediction.

We hypothesized that autistic children would be greatly affected by certain sensory stimulation. While TD children can quickly recover to a normal state after the sensory trial, autistic children may be slower to return to normal. In this paper, we describe our experiments and the ASD prediction models we developed based on the ECG and SC response signals recorded during the sensory trials.

II. Literature Review

Current ASD detection mechanisms are based on the observation of a subject's social interaction by either close observers or behavioral therapists [5]. The ASD assessment instruments are often lengthy, require extensive training before they can be administered, and are in general not very accurate [2].

Some researchers have argued that PsNS activity can be used as an indicator for the presence of autism and machine learning-based approaches can be utilized to build predictive models for its detection. Laufer and Nemeth [6] used SC to predict user action, based on a neural network model, by collecting SC data while users were playing an arcade game. Changchun et al. [7] designed a therapist-like support vector machine (SVM)-based affective model as part of a computer-based ASD intervention tool for children using physiological responses that predicts autism with an accuracy of 82.9%.

Much of the existing research in the field of time series analysis is relevant for this study. Dynamic time warping (DTW) [8] is a popular technique that can be used to compare two time-dependent series with different time deformations and speeds. For example, Muda et al. [9] used DTW to create efficient voice recognition algorithms. Juang [10] used DTW based hidden markov models and linear predictive coding techniques to develop speech recognition models. To optimize DTW, Salvador and Chan introduced FastDTW [11], which

* Corresponding Author

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_6

is an approximation of DTW with linear time and space complexity and is thus comparatively fast. Mueen et al. have introduced several variants of DTW, including constrained DTW, multidimensional DTW and asynchronous DTW [12].

Piecewise linear approximation (PLA) is one of the most common ways to process time series. It works by approximating a time series of length l with n straight lines using different algorithms, such as the top-down, bottom-up and sliding window approaches. Keogh at el. [13] developed a sliding window and bottom-up algorithm as a means to derive PLA and perform segmentation of time series. Some methods represent time series using motifs, which are derived by identifying frequently occurring patterns in the time series and replacing each pattern with a symbol. Lonardi et al. introduced an algorithm, called enumeration of motifs (EoM) [14], that uses matrix approximation to locate repeated patterns in the time series. Lin et al. introduced the symbolic aggregate approximation (SAX) [15] method, which discretizes original time series data into strings and defines distance measures on the symbolic string representation. Looking for a way to characterize co-evolution patterns in time series, Anastasiu et al. [16] devised an optimal segmentation algorithm that segments users' individual series into varying length segments represented by one of k patterns shared by all the users.

III. Dataset

Our research is based on examining existing data from a study conducted by Dr. Megan C. Chang [3]. The data were collected from various sensors during a SCP [17] in which the reactions to multiple stimuli were observed from 25 children with ASD and 25 typically developing (TD) children between 5 and 12 years of age. Each protocol took 45–90 minutes including preparation, and had three phases: baseline, sensory challenge, and recovery. The baseline and recovery periods lasted 3 minutes each and did not include any stimulation. The sensory challenge consisted of six different sensory stimuli with a pseudorandom pause of 12–17 seconds between the stimuli. Each stimulus was administered for 3 seconds and was presented at least 8 times. Following are the six stimuli, listed in the order they were administered:

- auditory – continuous sound tone of 84 decibels
- visual – 20W strobe light at 10Hz
- auditory – interrupted sound siren at 78 decibels
- olfactory – wintergreen oil passed under the nose
- tactile – touch along the jaw bone with a feather
- vestibular – chair tilted back to a 30 degree angle

Physiological ECG and SC data were continuously collected from multiple sensors in response to the eight stimuli (including the baseline and recovery periods). To obtain an index of PsNS function, ECG activity was collected by placing sensors on the child's chest. To measure the SNS activity, galvanic skin response was measured by attaching sensors to the right hand of the child. The sweat glands secrete more sweat as the subject becomes excited or nervous, which in turn increases skin conductance. The ECG and SC data were collected at a frequency of 500Hz and 40Hz, respectively. This resulted in a very long multivariate time series consisting of approximately 3 million correlated values across the length of the series. Table I provides a description of the dataset that was collected from the 50 subjects.

TABLE I
Dataset Description

# Autistic samples	25
# TD samples	25
Average # data points per subject	2,981,476
Average # data points per stimulus	372,682

Fig. 1 shows an example of the ECG and SC data for a subject spanning 10 seconds. The left y-axis shows the ECG signal, measured in milli-Volts (mV), and the right y-axis shows SC intensities, measured in micro-Siemens (μSiemens).

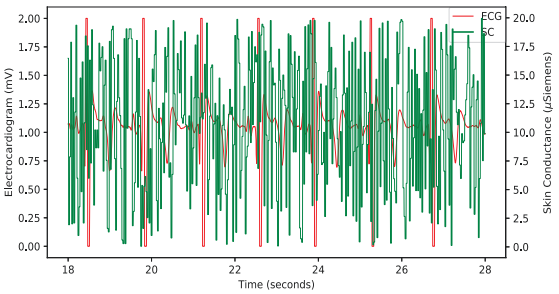


Fig. 1. Time series showing 10 seconds of ECG and SC signal for a subject. The figure is best viewed in color.

IV. Hypothesis and Supporting Evidence

We hypothesize that autistic children are greatly affected by certain sensory stimulation and thus may take longer to return to a normal state than TD children, who can quickly recover after the sensory trial. To test this, we compared the sensory data recorded during an initial baseline rest stage of the protocol, recorded prior to any stimulus being administered, with data recorded during the final recovery rest stage, 30 seconds after the final stimulus was administered. No stimulus was administered during either rest stage. For each subject, we compared the baseline and recovery rest stages by computing the Euclidean DTW distance of the ECG and SC time series recorded during the rest periods.

To analyze the differences between the baseline/recovery distances of autistic and TD children, we fit a Gaussian probability distribution function (PDF) over the distances between the baseline and recovery sensor time series data for autistic and TD children. Fig. 2 shows these functions for the ECG time series. Results show that autistic (solid green line) children exhibit substantially greater differences between their respective baseline and recovery phases than TD children (dashed red line). The PDF means for autistic and TD children were 1.25e+9 and 9.07e+8 and their standard deviations were 6.9e+8 and 4.03e+8, respectively. Results suggest that TD children recover faster, which would explain the shorter distances between the respective baseline and recovery phase time series.

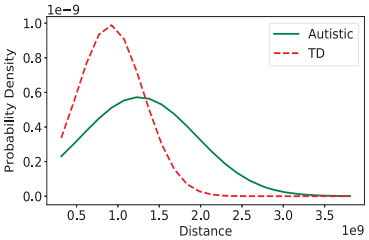
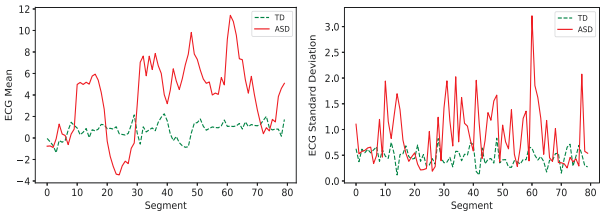


Fig. 2. ECG Gaussian probability density functions of DTW distances between the baseline and recovery stages for autistic and TD subjects.



(a) (b)

Fig. 3. Plot showing ECG mean (a) and standard deviation (b) values for a TD subject (dashed) and an autistic subject (solid).

V. Methods

In the remainder of the paper we describe predictive models we developed for autism detection from the ECG and SC response signals recorded during the sensory trials.

A. Feature Extraction

As a means to improve analysis efficiency, we propose to transform the data in a form that is representative of the input signal but has much smaller uniform dimensionality. We devised three different methods to extract features that can be used to conduct specific experiments.

1) Equal Width Partitioning (EWP): During the SCP, a particular stimulus is administered in a specific number of contiguous trials at equal intervals. Thus, we can divide the data into sub-series and still capture the patterns or trends in the series. In this approach, for each subject, the ECG and SC data were first split into 8 equal parts representing the 8 stimuli. The data were then standardized using the mean and standard deviation of the baseline stage, i.e., the first of the 8 splits, which captures the normal ECG and SC signal for a subject prior to any stimulus. The data for each stimulus were then split into n equal parts, and two different approaches were used to encode the information in each split and create different machine learning models for ASD prediction in children using either ECG data, SC data, or both data types.

a) Mean and standard deviation (MSD) representation: In this approach, we represented the n splits for each stimulus using the mean and standard deviation of the data in that split. The final data vector consists of n ECG mean and standard deviation values followed by n SC mean and standard deviation values for each stimulus. Fig. 3 shows the ECG mean and standard deviation values for a TD subject (dashed green line) and for an autistic subject (solid red line) chosen at random. One can observe that the ECG mean and standard deviation values of the autistic subject are generally higher than those of the TD subject. The maximum mean value for the autistic subject is 9.52 and that for the TD subject is 5.08.

b) Slope and intercept (SI) representation: We assume that an autistic child gets more excited when a stimulus is administered as compared to a TD child. When a subject gets excited or nervous, his/her ECG values spike, showing higher maximum and minimum values, and his/her sweat glands secrete more sweat, which in turn increases skin conductance. Thus, we hypothesize that the trend and intensity of the signal contains sensitive information that can be used to predict ASD.

For each of the n splits and for each stimulus, we retrieved all peak (maximum) values and all valleys (minimum) values of ECG data in a cycle. A data point is considered a peak/valley value if its value is greater/smaller than the value of its neighboring data points. After retrieving all peak and valley values in a time series, we represented each split as the slope and intercept of the *best fit line* (BFL) for both peak and valley values. SC values fluctuate less than ECG values do, in general. Therefore, we represented the n splits for each stimulus with the slope and intercept of the BFL over the entire SC time series data in that split. The slope of the BFL captures the variation in trend and the intercept captures the intensity of the signal.

Fig. 4 shows the valley-based slope and intercept representation of the ECG time series and slope and intercept representation for the SC time series, for a TD subject (dashed green line) and for a subject with ASD (solid red line), chosen at random. Time series data represented in these figures were processed using $n = 10$. One can observe that the variation in slopes, especially for ECG valley points and SC data, is higher for the autistic subject as compared to the TD subject. SC data shows more discriminatory characteristics, with autistic subjects showing higher maximum and minimum slope values. One can also observe that the intensity of the signals (ECG and SC), as shown by the intercept graphs, is much higher for autistic subjects as compared to TD subjects.

2) Dynamic Time Warping (DTW): The approach we devised in Section V-A1 transforms the real time series data into a derived format, which may lead to some loss of information. DTW allows us to compare two time series in their raw format. As DTW automatically accounts for time deformations, it will identify similar patterns in two time series even if one of them is longer than the other. In this approach, we used FastDTW, which is an approximation of DTW that has linear time and space complexity [11] to compare the ECG or SC time series between two subjects. Due to the very large size of our time series, both the original DTW and the FastDTW methods failed to compute the distance between our time series for different stimuli on on our very large server with 24 GB of random access memory (RAM), both running out of available memory. We thus split the series into 8 subsequences with r% overlap, since each stimulus was repeated 8 times, computed distances between the ith sub-sequence of

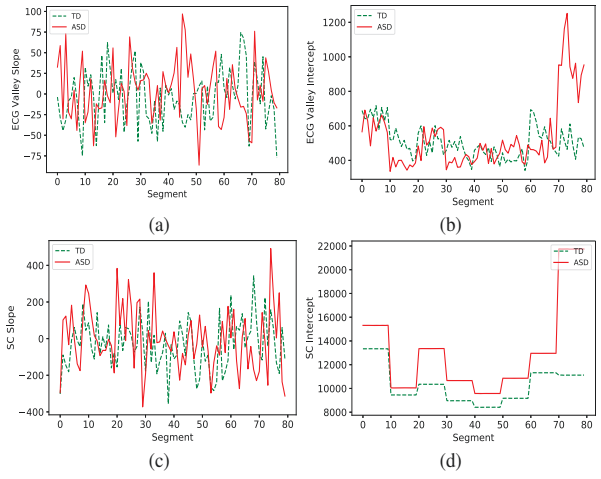


Fig. 4. Plot showing valley-based slope (a) and intercept (b) representation of the ECG time series and slope (c) and intercept (d) representation of the SC time series for a TD subject (dashed) and an autistic subject (solid).

the candidate sequences, and used the maximum of these subsequence distances as the distance between the two candidate sequences. We also tried the efficient DTW method introduced by Mueen et al. [12] and compared it with FastDTW. While it was marginally faster than FastDTW, it required more memory and most of our series could not be computed on our server due to lack of available memory.

3) Symbolic Representation of Time Series: In this approach, we used SAX [15] to represent each of the time series using a SAX vector with a given number of symbols and segments. To get the best representation, we tested with numbers of symbols in the range 2 to 10 and numbers of segments from 2 to 14, in increments of 1. After representing the time series using SAX, we computed pairwise Euclidean DTW distances. These distances were then used to create a KNN-based ASD prediction model.

B. Developing Prediction Models for Autism Detection

1) Base Models: In our experiments, we aim to classify the subject as either Autistic or TD. To perform this binary classification, we trained and tested models using the following methods:

- decision tree (DT)
- k-nearest neighbor (KNN)
- support vector machine (SVM)
- naïve Bayes (NB)
- random forest (RF)
- XGBoost (XGB)
- DTW-based KNN (DTWNN)
- SAX-based KNN (SAXNN)

The first six models consume the features generated using methods specified in Section V-A1. Separate models were created using the MSD and SI feature generation approaches. The DTWNN model is based on the method described in Section V-A2, which utilizes the raw time series for comparison and prediction. The SAXNN model is based on the method described in Section V-A3, which first transforms the raw time series data into its SAX representation before computing pairwise Euclidean DTW distances between the subjects. As we have both ECG and SC data, we wanted to understand how different physiological data help in predicting autism. Thus, we created different models using either only ECG data, SC data, or both ECG and SC data.

2) Ensemble Models: In Section V-B1, we executed experiments for each separate stimulus. After building the separate models for all stimuli, we combined them to build ensemble models and make additional predictions. We used three different approaches to create ensemble models.

a) Majority vote: In this approach, we combined the predictions from all the models for different stimuli and chose the majority predicted class as the final prediction. All the model outputs were given the same weight.

b) Weighted prediction: In this approach, instead of giving the same weight to all the model outputs, we weighed the classification output of each stimulus model with the prediction confidence of its associated model, which ranges between 0 and 1. Considering a vector $\mathbf{w}_c$ of weights associated with each stimulus and the vector $\mathbf{y}$ representing classification predictions of models associated with each stimulus, we compute the final prediction as the linear combination of vectors $\mathbf{w}_c$ and $\mathbf{y}$, $y_c = \mathbf{w}_c{}^T\mathbf{y}$. The vector $\mathbf{y}$ contains the predicted classes, +1 or -1, representing TD and autistic subjects, respectively. A negative y_c prediction value indicates that the models predicted the subject as autistic with higher confidence.

c) Stochastic gradient descent (SGD): In this approach, instead of using the prediction confidence scores from separate stimuli models as weights, as described in Section V-B2b, we learned the contribution of each stimulus towards predicting autism. Some stimuli may contribute positively towards correct prediction, while others may contribute negatively. This can be done by deriving a set of weights such that the linear combination of the weight vector and predictions from different stimulus models results in an accurate binary classification of autistic and TD children. The weight vector $\mathbf{w}_s$, is learned via the SGD algorithm applied to training set predictions. Then, the stimuli predictions in the test set are combined linearly with the weights to generate the final SGD predictions for test samples, computed as $y_s = \mathbf{w}_s{}^T\mathbf{y}_s$.

VI. Experiment Design

We used *accuracy* as the performance measure when comparing the prediction models. Accuracy is an appropriate evaluation metric in our setting, as the dataset contains an equal number of samples for both autistic and TD subjects. It is defined as

$$A = \frac{T_p}{T_s} \times 100,$$

where T_p represents the total number of correct predictions and T_s represents the total number of subjects.

We measure efficiency as the training and prediction runtime, in seconds, for each of the different models. Prediction time is given priority over training time, as training can be

done offline but prediction must be executed online, in real time, and thus needs to be fast.

For each prediction or time series analysis method we tested, we tuned available hyper-parameters to obtain the highest possible effectiveness using that method. Due to lack of space, the details of the hyper-parameter tuning can be found in [18].

VII. Results And Discussion

A. Effectiveness Results

1) Base Models: We created eight different models, as described in Section V-B, one for each of the eight stimuli. The first six models, namely, DT, KNN, SVM, NB, RF and XGB, were built using the features extracted based on the two approaches mentioned in the EWP method described in Section V-A1, which splits the time series into a specified number of sections. We created different dataset representations with number of splits, n, ranging from 2 to 13, inclusive. For each value of n, after further splitting the training set into training and validation subsets, we trained different instances of all the six models using different combinations of hyper-parameters. Then, we chose the best model instance based on its validation accuracy. Finally, we re-trained the best model for each algorithm using the chosen best hyper-parameters and the entire original training set.

The DTWNN model utilizes the features extracted using the DTW approach mentioned in Section V-A2, which computes the Euclidean DTW distance between different subjects. Higher distance values imply lower similarity, and *vice versa*. For creating the overlapping splits, we chose r = 10%. The SAXNN model was then built using the SAX feature construction method described in Section V-A3.

Fig. 5 (a) shows the comparison of the best base performing model instances for different algorithms, created using different feature extraction methods and using baseline stage data. We observed that, in almost all cases, the models created using SI features perform better than those created using MSD features. Also, among the two standard time series approaches, the models created using SAX features perform much better as compared to those based on DTW distances.

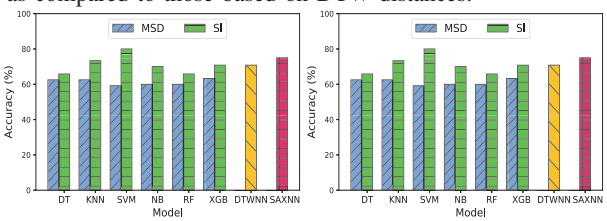


Fig. 5. (a) Comparison of the best base models for the auditory (tones) stage. (b) Comparison of the best SGD ensemble models.

Table II shows the accuracy scores of the best models for each stimulus. Auditory (tones) and visual stimuli data result in the best performing models, with an accuracy of 80.00% (highlighted in bold). We also observed that two of the best performing models utilize both ECG and SC data for making predictions, showing that both types of sensor data are important in predicting autism.

TABLE II
Best Base Model Accuracy Values Using Each Stimulus

	Accuracy(%)	Model	Data Used
Baseline	75.83	SAXNN	SC
Auditory (Tones)	**80.00**	SVM	Both
Visual	**80.00**	XGB	SC
Auditory (Siren)	77.50	RF	ECG
Olfactory	77.50	SAXNN	SC
Tactile	74.17	SAXNN	SC
Vestibular	78.33	RF	Both
Recovery	73.33	SAXNN	Both

2) Ensemble Models: We combined the results from the models generated using different stimuli, presented in Section VII-A1, to create ensemble models. We compared the accuracy of the ensemble models with the best base models. Ensemble models were created using the three approaches described in Section V-B2.

Fig. 5 (b) shows the comparison of the best SGD ensemble models. We observed that models constructed from SI features outperformed those using MSD ones in almost all cases. The best performing model using SI features is an SGD ensemble XGB model that achieved an accuracy of 93.33%, which is 7.50% higher than the best performing model using MSD features.

As SI features performed better than the MSD ones, further comparisons with DTW and SAX-based approaches were done using only SI features. As mentioned in Sections V-A2 and V-A3, both DTW and SAX-based models are KNN models. Table III shows the best model accuracies for the different tested data processing and modeling methods. One can observe that all the models give the best accuracy using the SGD ensemble method. In this ensemble approach, as described in Section V-B2c, the SGD algorithm is applied on the training set to learn the weights of each stimulus towards making correct predictions.

TABLE III
Best Ensemble Model Accuracy Values

	Accuracy(%)	Ensemble Type	Data Used
DT	92.50	SGD	Both
KNN	81.67	SGD	SC
SVM	87.50	SGD	Both
NB	88.33	SGD	SC
RF	89.17	SGD	Both
XGB	**93.33**	**SGD**	**Both**
DTWNN	77.50	SGD	Both
SAXNN	92.50	SGD	ECG

The best overall performing model was the SGD ensemble XGB model, built using both ECG and SC data, which resulted in an accuracy of 93.33%. The value is approximately 4.16% greater than that achieved using either the majority vote or weighted prediction vote ensemble methods.

As the best accuracy is achieved using both ECG and SC data, we can infer that both types of sensors are important in accurately predicting autism. Additionally, we observed that the next best performing models were DT and SAXNN,

which were built using either only ECG data or both ECG and SC data. This further highlights the importance of ECG data in predicting autism in children. In comparison to the best performing base model, the ensemble models performed much better in general. The best performing ensemble model (93.33%) had an accuracy that was 13.33% higher than the best performing base model (80.00%). Even ensemble models built using majority vote (89.17%) and weighted prediction (89.17%) decisions performed better than the base models.

Even though DTW is an important metric for comparing time series, we observed that classification models based on DTW failed to outperform other classification models in our problem. The best accuracy achieved by the DTWNN models was 77.50%, which is approximately 18% lower than that of the best performing model.

B. *Efficiency Results*

We measured the efficiency of the models based on the time taken to train and perform predictions. Fig. 6 shows the comparison of natural log transformed training and prediction times, in seconds. The log scaling in the figure is necessary due to the very wide range of values, which would otherwise hide most results in the graph.

The best performing model in terms of accuracy was the XGB model, which was the third slowest method, taking approximately 49,300 seconds to train and 1.23e−4 seconds to predict. On the other hand, the DTW-based model took approximately 4.40 times longer to train and 10^8 times longer to predict in comparison to the SAXNN model. The high execution time for training and prediction makes it difficult to utilize DTW-based models in real-world applications for our problem. On the other hand, the DT model achieved the second highest accuracy (92.50%) and predicts 7 times faster than the best performing XGB model.

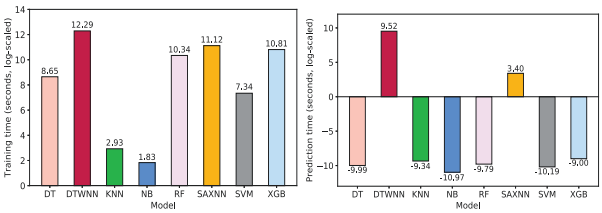


Fig. 6. Comparison of training time (left) and prediction time (right) for all methods.

VIII. Conclusions

In this paper, we described novel techniques we developed for analyzing very large time series of ECG and SC sensor data derived from a sensory trial administered to 50 autistic and TD children. Our analysis showed that autistic children are affected to a higher degree by some stimuli as compared to TD children and take longer to recover. Moreover, the feature extraction methods we developed were both effective and efficient in analyzing multivariate time series with over 2 million values. An XGB-based model trained on vectors constructed using a feature engineering method we developed (SI) achieved the best performance (93.33% accuracy) taking only a millisecond to predict samples.

References

[1] M. Norris and L. Lecavalier, "Screening accuracy of level 2 autism spectrum disorder rating scales: A review of selected instruments," *Autism*, vol. 14, no. 4, pp. 263–284, 2010, doi:10.1177/1362361309348071.

[2] J. Constantino and C. Gruber, "Social responsive scale (srs) manual," *Los Angeles, CA: Western Psychological Services*, 2005, doi:10.1177/1534508410380134.

[3] M. C. Chang, L. D. Parham, E. I. Blanche, A. Schell, C.-P. Chou, M. Dawson, and F. Clark, "Autonomic and behavioral responses of children with autism to auditory stimuli," *American Journal of Occupational Therapy*, vol. 66, no. 5, pp. 567–576, 2012, doi:10.5014/ajot.2012.004242.

[4] T. Chaspari, M. Goodwin, O. Wilder-Smith, A. Gulsrud, C. A. Mucchetti, C. Kasari, and S. Narayanan, "A non-homogeneous poisson process model of skin conductance responses integrated with observed regulatory behaviors for autism intervention," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1611–1615, doi:10.1109/ICASSP.2014.6853870.

[5] S. Chandler, T. Charman, G. Baird, E. Simonoff, T. Loucas, D. Meldrum, M. Scott, and A. Pickles, "Validation of the social communication questionnaire in a population cohort of children with autism spectrum disorders," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 46, no. 10, pp. 1324–1332, 2007, doi:10.1097/chi.0b013e31812f7d8d.

[6] L. Laufer and B. Németh, "Predicting user action from skin conductance," in *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 2008, pp. 357–360, doi:10.1145/1378773.1378829.

[7] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Physiology-based affect recognition for computer-assisted intervention of children with autism spectrum disorder," *International journal of human-computer studies*, vol. 66, no. 9, pp. 662–677, 2008, doi:10.1016/j.ijhsc.2008.04.003.

[8] M. Müller, "Dynamic time warping," *Information retrieval for music and motion*, pp. 69–84, 2007, doi:10.1007/978-3-540-74048-3_4.

[9] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," *arXiv preprint arXiv:1003.4083*, 2010. [Online]. Available: http://arxiv.org/abs/1003.4083

[10] B.-H. Juang, "On the hidden markov model and dynamic time warping for speech recognitiona unified view," *Bell Labs Technical Journal*, vol. 63, no. 7, pp. 1213–1243, 1984, doi:10.1002/j.1538-7305.1984.tb00034.x.

[11] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007. [Online]. Available: https://content.iospress.com/articles/intelligent-data-analysis/ida00303

[12] A. Mueen and E. Keogh, "Extracting optimal performance from dynamic time warping," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 2129–2130.

[13] E. Keogh, S. Chu, D. Hart, and M. Pazzani, "Segmenting time series: A survey and novel approach," in *Data mining in time series databases*. World Scientific, 2004, pp. 1–21.

[14] J. Lonardi and P. Patel, "Finding motifs in time series," in *Proc. of the 2nd Workshop on Temporal Data Mining*, 2002, pp. 53–68.

[15] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and knowledge discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[16] D. Anastasiu, A. Rashid, A. Tagarelli, and G. Karypis, "Understanding computer usage evolution," in *2015 IEEE 31st International Conference on Data Engineering, ICDE 2015*, vol. 2015-May. IEEE Computer Society, 5 2015. doi: 10.1109/ICDE.2015.7113424 pp. 1549–1560.

[17] R. C. Schaaf, T. W. Benevides, E. Blanche, B. A. Brett-Green, J. Burke, E. Cohn, J. Koomar, S. J. Lane, L. J. Miller, T. A. May-Benson *et al.*, "Parasympathetic functions in children with sensory processing disorder," *Frontiers in Integrative Neuroscience*, vol. 4, p. 4, 2010, doi:10.3389/fnint.2010.00004.

[18] M. Kapoor and D. C. Anastasiu, "A data-driven approach for detecting autism spectrum disorders," San José State University, San José, CA, USA, Tech. Rep. 2019-1, 2019. [Online]. Available: pdf/papers/2019-Kapoor-asp-tr.pdf

Optimal Regression Tree Models through Mixed Integer Programming*

Ioannis Gkioulekas
University College London)
Torrington Place, London, UK
ioannis.gkioulekas.16@ucl.ac.uk

Lazaros G. Papageorgiou
University College London)
Torrington Place, London, UK
l.papageorgiou@ucl.ac.uk

***Abstract*—Regression analysis is a tool for predicting the output variables from a set of known independent variables. Through regression, a function that captures the relationship between the variables is fitted to the data. Tree regression models are popular in the literature due to their ability to be computed quickly and their simple interpretations. However, creating complex tree structures can lead to overfitting the training data resulting in a poor predictive model. This work introduces a tree regression algorithm that employs mathematical programming to optimally split data into two sub regions, called nodes, and a statistical test to assess the quality of partitioning. A number of publicly available literature examples have been used to test the performance of the method against others that are available in the literature.**

***Index Terms*—Mathematical programming, Regression analysis, Decision trees, Optimization**

I. Introduction

Regression analysis is a predictive modeling technique for formulating the correlation between a set of dependent and independent variables. Depending on how the values of the independent variables (i.e. predictors) vary, the value of the dependent variable (i.e. response) also changes. The objective of regression is to apply a mathematical function to the data that captures these changes.

Examples in the literature include linear regression, Support Vector Machine Regression (SVM) [1], K-nearest neighbors (KNN) [2], Multivariate Adaptive Regression Splines (MARS) [3] and Random Forest [4].

In the field of mathematical programming there is the Automated Learning of Algebraic Models for Optimization (ALAMO) [5], [6] and two piecewise regression approaches called Optimal Piecewise Linear Regression Analysis (OPLRA) [7] and Piecewise Regression with Optimised Akaike Information Criterion (PROA) [8], [9].

In every machine learning application, overfitting is a problem that needs to be addressed. Overfitting is the result of describing a set of training data but also its noise. The resulting model is designed to perfectly explain this specific set of data without capturing the 'true' mechanism that generates them, leading to a very poor predictive performance. On the opposite end there is underfitting, the process of constructing a very simple model that is not capable of capturing the information that exists in the data [10].

A. Contribution of this work

The focus of this work is generating tree regression models. Even though the simplicity of linear models cannot be overlooked, their predictive accuracy is often lacking. By using piecewise approaches, it is possible to partition the data into multiple subsets and fit linear functions to each one. This way, the accuracy of the overall model is greatly improved, while still maintaining model simplicity.

However, identifying the splitting points can be a complex task involving a lot of trial and error. The segmented package, which is part of the R library [11], is able to perform piecewise analysis [12], [13]. This package fits segmented regression expressions to a set of data, but the user has to specify the number of regions as well as estimates for the position of the break points. The method then iterates until the final break points have been identified.

The OPLRA and PROA methods that were mentioned earlier, are mathematical programming based methods that can optimally partition the data into multiple regions by automatically selecting the value of the break points. However, both methods can only identify a single partitioning variable, with the latter using the Akaike Information Criterion to decide the optimal number of partitions. However, tree regression approaches have the advantage of partitioning the input space on more than one variables if necessary.

In this work, a mathematical programming based approach is proposed that is able to generate tree regression models. This novel approach requires a numerical multivariate dataset as input and then, by using an optimization model and a post-processing statistical test, can generate tree structures that minimise the mean absolute error (MAE) of the fitting.

By accommodating statistical testing as a post-processing step, the algorithm can assess whether or not the partitioning of the data is meaningful. This way, an established criterion is used to control the tree generation process, hence address-

* This work was supported by the Leverhulme Trust under Grant number RPG-2015-240

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_7

ing the issue of having a good balance between predictive performance and model complexity.

II. Mathematical Formulation

A. Optimal model for partitioning

In this section, the mathematical programming model that was used in the `MPtree` algorithm is described, as formulated by [14] in the literature. This model is responsible for partitioning a node into two child nodes based on a particular partitioning feature. The mathematical model is presented is as follows:

Indices

c	child node of the current parent node; $c = left$ represents left child node, and $c = right$ represents right child node
m	feature/independent input variable
m^*	the feature where sample partitioning takes place
n	the current node
s	data samples, $s = 1, 2, ..., S$

Sets

C_n	set of child nodes of the current parent node n
S_n	set of samples in the current parent node n

Parameters

a_{sm}	numeric value of sample s on feature m
y_s	output value of sample s
u_1, u_2	arbitrary large positive numbers
ϵ	very small number

Continuous variables

B_c	intercept of regression function in child node c
D_s	training error between predicted output and real output for sample s
Pr_{sc}	predicted output for sample s in child node c
$W1_{mc}$	regression coefficient for feature m in child node c
$W2_{mc}$	regression coefficient for feature m in child node c
X_{m^*}	break-point on partitioning feature m^*

Binary variables

F_{sc}	1 if sample s falls into child node c; 0 otherwise

Mathematical Constraints

In order to assign samples into the child nodes, binary variables are introduced to the model in the following constraints:

$$a_{sm} \leq X_m + u_1 \cdot (1 - F_{sc}) - \epsilon \quad \forall\ s \in S_n,\ c = left,\ m = m^* \tag{1}$$

$$X_m - u_1 \cdot (1 - F_{sc}) + \epsilon \leq a_{sm} \quad \forall\ s \in S_n,\ c = right,\ m = m^* \tag{2}$$

The following constraint restricts that each sample belongs to only one child node:

$$\sum_{c \in C_n} F_{sc} = 1 \qquad \forall s \in S_n \tag{3}$$

For each child node c, polynomial functions of order 2 are employed to predict the value of samples (P_{sc}):

$$P_{sc} = \sum_m a_{sm}^2 \cdot W2_{mc} + \sum_m a_{sm} \cdot W1_{mc} + B_c \quad \forall\ s \in S_n,\ c \in C_n \tag{4}$$

For any sample s, its training error is equal to the absolute deviation between the real output and the predicted output for the child node c where it belongs to and can be expressed with the following two equations:

$$D_s \geq y_s - P_{sc} - u_2 \cdot (1 - F_{sc}) \quad \forall\ s \in S_n,\ c \in C_n \tag{5}$$

$$D_s \geq P_{sc} - y_s - u_2 \cdot (1 - F_{sc}) \quad \forall\ s \in S_n,\ c \in C_n \tag{6}$$

The objective function is to minimise the sum of absolute training errors of splitting the current node n into its child nodes:

$$min \sum_{s \in S_n} D_s \tag{7}$$

The resulting model can be summarised as:
objective function (7)
subject to (1)-(6) constraints
and is formulated as an MILP problem that can be solved to optimality.

B. The Chow statistical test

In regression, the F statistical test can be used to assess the quality of stepwise regression. By assuming that there is a break in a dataset and splitting it into two subsets, a regression model can be applied to each subset, resulting in a better fit compared. However, such an approach will add to the complexity of the model and might lead to overfitting. Therefore, the Chow test can be applied to compare the predictive performance of a segmented and a non-segmented regression model.

Suppose that there are two subsets and the question is whether to perform regression on the entire dataset consisting of both subsets (we denote this model 1), or to apply separate regression models for each subset (we denote this model 2). So RSS_1 is the residual sum of squares for the first model and RSS_2 is the residual sum of squares for model 2 (which in this case is the sum of the RSS for each subset). In general, there will be an improvement when splitting the data ($RSS_2 \leq RSS_1$), with equality occurring only when the all regression coefficients for the two models coincide [15]. However, there is a trade-off due to the added complexity of the overall regression model. By splitting the data into two subsets and performing separate regressions, more parameters

are added to the model and hence more degrees of freedom. So the chow is useful for testing if there is a statistically significant difference in predictive performance.

The F statistic for the chow test can be computed as follows [15]:

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{p_2 - p1}\right)}{\frac{RSS_2}{n - p_2}} \tag{8}$$

where:

RSS_1	Residual sum of squares of model 1 (single regression for the entire dataset)
RSS_2	Total residual sum of squares of model 2 (separate regression for each subset)
p_1	The regression parameters of model 1
p_2	The regression parameters of model 2
n	The total number of samples in the dataset

The null hypothesis states that model 2 **does not** provide a significantly better fit than model 1. So the procedure to either reject or accept the null hypothesis is as follows:

- Calculate the F statistic using equation (8)
- Choose an appropriate confidence level (e.g. 99%)
- Calculate the critical F_{crit} value of the F-distribution
- Reject the null hypothesis if $F > F_{crit}$

According to the steps above, if there is evidence to reject the null hypothesis, it is accepted that model 2 **does** provide a significant improvement in predictive performance.

C. Application to tree regression

The use of this test can aid the process of generating regression trees, since it can be used as a criterion for splitting nodes. Every node can be considered a population that can be split into two separate subsets and a decision has to be made of either splitting the data into two child nodes or stop the tree generation process (in other words apply a single regression model to the node).

The algorithm for building the tree is given below. As with other tree algorithms, recursive splitting is used to generate the tree. For each node, the optimization partitioning model is applied to split the node into two child nodes. Then the Chow test is applied comparing the linear regression model with the segmented regression model. If there is a significantly better predictive performance by splitting the node, then the partitioning is approved and the algorithm starts again by following the same procedure for all the new nodes. However, if a node splitting is rejected then this node will no longer be consider for splitting. The entire tree generation process is terminated when there are no more nodes that are eligible for splitting.

In the previous work, the proposed `MPtree` algorithm used a heuristic approach to control the tree generation process. This heuristic introduced a new parameter which was used as a threshold to the reduction percentage of the absolute deviation. Additionally, the heuristic used this parameter to compare the splitting of each node with the original error in the root node. By performing a sensitivity analysis, that parameter was set to the value of 0.015.

The proposed algorithm, from now on called `StatTree`, is briefly explained below.

StatTree algorithm	
Step 1.	Select a node as the current node and fit a linear regression model of order 2 minimising absolute deviation. Post process the results and calculate $RSS^{current}$.
Step 2.	In each current root, for each input variable m, specify it as splitting variable ($m = m^*$) and solve the *MPtree* model. The absolute deviation is noted as $ERROR_m^{split}$.
Step 3.	Identify the best split corresponding to the minimum absolute deviation, noted as $ERROR^{split} = \min_m ERROR_m^{split}$.
Step 4.	Post process the results and for the best partitioning feature calculate RSS^{split}.
Step 6.	Calculate the F statistic using equation 8
Step 7.	Choose a confidence level of 99% ($\alpha = 0.01$)
Step 8.	Calculate the F_{crit} value of the F-distribution. If $F > F_{crit}$, then the current node is split; otherwise, the current node becomes a terminal node.

III. Computational part

A. Examined datasets

A number of examples are considered in this report, all of which are summarised in table (I) that follows. Those datasets are derived from different online sources. More specifically the pharmacokinetics, earthquake, abalone and speeding datasets are available through a package in **R**, boston, bodyfat and sensory datasets are available through **StatLib** [16], concrete, cooling, heating, wine and yacht through the **UCI** machine learning repository [17] and the rest are available through the **KEEL** dataset repository [18].

TABLE I
Regression datasets examined in this work

Data	Predictors	Size	Data	Predictors	Size
Concrete	8	1030	Octane	4	82
Cooling	8	768	Pharma	4	132
Heating	8	768	Plastic	2	1650
Yacht	6	308	Sensory	11	576
Bodyfat	14	252	Wankara	9	1609
Boston	13	506	Abalone	8	4177
Dee	6	365	Speeding	3	8437
Earthquake	4	1000			

The `Yacht` hydrodynamics set predicts the residuary resistance of sailing yachts for evaluating the ships' performance and for estimating the required propulsive power. An assessment of heating and cooling load requirements is captured

in the Energy Efficiency dataset [19], of different buildings as a function of 8 parameters. The Concrete dataset [20] to predicts the compressive strength of concrete as a structural material.

A study of the kinetics of the anti-asthmatic drug theophylline is included in the Pharma dataset. Twelve subjects were given oral doses of the drug and the aim is to predict the final theophylline concentration of each subject by measuring parameters such as weight and time. Earthquake data based on the location of seismic events that occurred near Fiji since 1964 are in the earthquake dataset. The Bodyfat dataset uses features such as age, weight and height to measure the percentage of bodyfat in a subject. An evaluation of wine quality by 6 judges is recorded in the Sensory dataset.

Dee, predicts the daily average price of electricity in Spain. The dataset contains values about the daily consumption of energy from various sources such as hydroelectric, fuel, natural gas and more. Plastic, computes how much pressure can a given piece of plastic withstand when a force is applied on it at a fixed temperature. Wankara, contains observations about weather information of Ankara during 1994-1998, with the goal of predicting the average temperature. Abalone, predicts the age of abalone from physical measurements which are easy obtain. The Speeding dataset has been collected from a study that tried to identify the effect of warnings signs on speeding patterns. The speed measurements were taken before the erection of a warning sign, after shortly after the erection of the sign and finally after the sign had been in place for some time. Finally, Boston consists of observations that predict the price of houses in various places in Boston.

B. Validation of the method

The simplest way to evaluate a model is to split the original data into two subsets, one for training and one for testing. The training set will be used to construct a regression model, which will be evaluated by using the testing set. The reason for doing so is to measure how well the model generalises to new, previously unseen data [21].

Cross-validation is a statistical method of evaluating the performance of models that is more reliable than simply splitting the data into two sets. The most common form of cross-validation is *k-fold* where the data is split into k subsets of equal size. Then the method uses one of these sets for testing and the rest for training. The method stops when all of the k sets have been used as the testing set. Parameter k is user-specified and is usually set to either 5 or 10 [21].

In this work, 5-fold cross-validation is selected to evaluate the performance of the proposed algorithm. 10 runs will be performed and the Mean Absolute Error (MAE) between model prediction and the true data will be calculated for each fold. The final score is the average of all the runs. The mathematical programming model that is responsible for bipartitioning the nodes, is implemented in the General Algebraic Modeling System (GAMS) [22] and solved using the CPLEX solver with optimality gap set at 0 and a time limit of 200s. The R programming language [11] is used for the *k fold* cross-validation procedure. The caret package [23] that is available in R, contains tools for data splitting, pre-processing, feature selection and more. In this work, the package is used to create random partitions of the samples and perform *k fold* to evaluate the predictive accuracy of the methods.

A number of tree methods from literature are also implemented in this work for comparison purposes on the same datasets. The methods include M5P regression [24], [25], CART regression [26], Cubist [27] and MPtree. All of those methods are implemented in the R programming language using the RWeka [28], rpart [29] and Cubist [30] packages respectively. R and GAMS were used for the MPtree algorithm. The same 10 runs of 5-fold cross-validation are performed to evaluate and compare with the proposed methods.

IV. Results

A. Comparison of the algorithms

As a pre-processing sterp, *feature scaling* is applied to each dataset according to the following equation:

$$\frac{A_{s,m} - min_s A_{s,m}}{max_s A_{s,m} - min_s A_{s,m}}$$

This is common practice since it prevents some variables from being more dominant than others in the final regression model.

TABLE II
Cross-validation results using MAE

	StatTree	MPtree	Cubist	CART	M5P
Concrete	4.344	4.868	**4.267**	7.239	4.656
Cooling	1.169	**0.891**	0.938	2.400	1.210
Heating	0.373	0.354	**0.347**	2.011	0.693
Yacht	**0.519**	0.539	0.557	1.669	0.931
Bodyfat	**0.183**	5.282	0.205	1.356	0.373
Boston	2.565	4.644	2.587	3.234	**2.501**
Dee	**0.313**	0.975	0.316	0.381	0.316
Earthquake	7.363	12.427	7.294	8.223	**7.273**
Octane	0.391	0.805	**0.384**	0.602	0.464
Pharma	0.908	**0.870**	1.053	1.339	1.328
Plastic	**1.227**	1.230	1.229	1.658	1.234
Sensory	0.607	0.663	0.602	**0.578**	0.601
Wankara	**0.968**	3.605	1.000	3.213	0.977
Abalone	**1.490**	1.512	1.500	1.731	1.521
Speeding	**4.171**	4.243	4.188	4.524	4.239

Table II contains the MAE results of all the runs of cross validation. For each dataset, the method that performed the best is marked with bold. StatTree has the best performance in terms of MAE score for 7 out of 15 examples. Cubist is the next best performer with 3 out 15, MPtree and M5P 2 out 15 and CART with only a single dataset. However, that alone is not a good indication of overall performance.

Constructing a figure to visualise the comparison of the various methods will aid the interpretation of the overall

predictive performance. In this figure, for each dataset the best performer is awarded 10 points whereas the worst performer is awarded 1 point. The final ranking is the average score that each method achieves across all datasets.

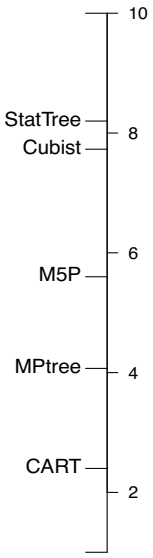


Fig. 1. Visualisation of the performance of the methods based on the MAE results

Looking at Fig. 1 it is easier to compare the overall performance of the methods. We can see that the `StatTree` algorithm is ranked at number 1. Also, a large performance gap exists between `StatTree` and `MPtree`, which indicates that the new proposed method is actually a better alternative. `Cubist` on the other hand, is the only method that can provide competitive results. However, since the performance of those two methods is very close, a statistical test has to be applied in order to check whether there is a significant difference in the results.

B. Statistical analysis

At this stage, the Welch's t-test can be applied. This is a two-sample test which is used to test the hypothesis that two populations have equal means and is reliable when the samples have unequal variances [31]. If we have evidence to reject this hypothesis using that test, then we can conclude that the difference between the two means is significant.

For each dataset, the two different populations that will be compared are the values of the 10 cross validation runs between the `StatTree` algorithm and one of the rest. If by performing the Welch's t-test there is evidence to reject the null hypothesis, then it can be concluded that there is a statistical significance between the two sample means, and the best method is the one that has the minimum average error.

The calculation of the t-statistic requires as input the average error values and the the variances of the cross validation runs. The next step is to calculate the p-values of a t distribution, which can be done through software, and test if there is evidence to reject the null hypothesis.

Because this is a two sample test, the analyses will occur in a pairwise fashion between `StatTree` and the other methods. By setting a significance level of $\alpha = 0.01$, the condition that has to be satisfied is the the probability values from the t-test $p < \alpha$. In this case the null hypothesis is rejected, which means that there is a difference between the two samples.

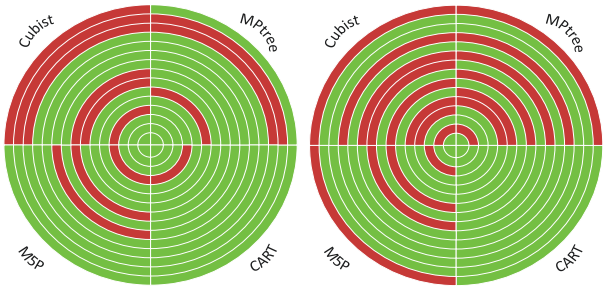


Fig. 2. Visualisation of the Welch's t-test. The circle on the left represents winner in terms of the MAE score. The circle on the right represents if there is a significant difference according to the t-test.

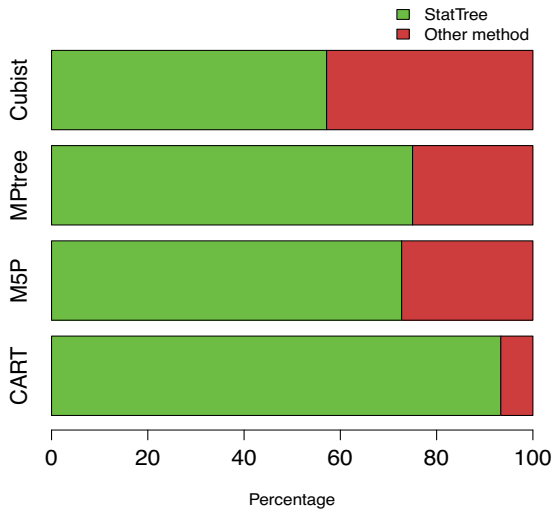


Fig. 3. Percentage of winning between `StatTree` and the other methods, based only on the datasets with meaningful statistical difference.

Fig. 2 is a visualisation of the statistical analysis. In this figure, each concentric circle represents a dataset, following the same order as in table II with `Concrete` being the outer ring and `Speeding` being the inner ring. The circles on the left compare the MAE scores of table II whereas the

circles on the right illustrate if there is a significant statistical difference between StatTree and the other methods. Green colour represents an advantage for StatTree, either better MAE score or statistical difference, whereas red the opposite. So, it is desirable that for each dataset the equivalent section in both circles to be green, because in that case StatTree has statistically better MAE score than the other examined method.

Fig. 2 and 3 summarise the findings of this work. Fig. 3 plots the percentage of the datasets for which StatTree is able to outperform the other examined approaches, but only for those that there is a statistically significant difference. For example, when compared to `CART`, there is a statistical difference between the MAE scores for all 15 examples and `StatTree` is outperforming `CART` in 14 out of 15. So that results in a 93% winning percentage. Accordingly, compared to `Cubist` the score 4 out of 7, compared to `M5P` the score is 8 out of 11 and finally compared to `MPtree` the score is 6 out of 8.

V. Conclusions

This work addresses the issue of multivariate regression analysis by generating tree structures. The proposed method uses an optimization model to split data into nodes and the Chow statistical test to avoid overfitting. The final algorithm is able to generate optimal tree structures by optimally deciding the partitioning variable and the value of the break point for every node and fits a regression model to each one.

To test the algorithm, several real world examples have been employed. The performance of the proposed method is compared to other established tree regression approaches from the literature. Computational experiments indicate that the new proposed method can consistently provide competitive results when compared to established tree regression methods, for a range of applications. Also, when compared to MPtree there is a big gain in predictive performance by utilising the Chow statistical test. Overall, Fig. 2 and 3 summarise the findings of this work, that the proposed approach is a competitive, and in some cases better, tree regression alternative to other established literature methods.

References

[1] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[2] K. T. Korhonen and A. Kangas, "Application of nearest-neighbour regression for generalizing sample tree information," *Scandinavian Journal of Forest Research*, vol. 12, pp. 97–101, 1997.

[3] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, vol. 19, pp. 1–67, 1991.

[4] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[5] A. Cozad, N. V. Sahinidis, and D. C. Miller, "Learning surrogate models for simulation-based optimization," *AIChE Journal*, vol. 60, pp. 2211–2227, 2014.

[6] Z. T. Wilson and N. V. Sahinidis, "The alamo approach to machine learning," *Computers & Chemical Engineering*, vol. 106, pp. 785–795, 2017.

[7] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "Mathematical programming for piecewise linear regression analysis," *Expert Systems with Applications*, vol. 44, pp. 156–167, 2016.

[8] I. Gkioulekas and L. G. Papageorgiou, "Piecewise regression analysis through information criteria using mathematical programming," *Expert Systems with Applications*, vol. 121, pp. 362–372, 2019.

[9] I. Gkiouleka and L. G. Papageorgiou, "Piecewise regression through the akaike information criterion using mathematical programming," *IFAC-PapersOnLine*, vol. 51, no. 15, pp. 730–735, 2018.

[10] D. M. Hawkins, "The problem of overfitting," *Journal of Chemical Information and Computer Sciences*, vol. 44, pp. 1–12, 2004.

[11] R Development Core Team, *R: A Language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2016. [Online]. Available: http://www.R-project.org

[12] V. M. R. Muggeo, "Segmented:An R package to fit regression models with broken-line relationships," *R news*, vol. 8, pp. 20–25, 2008.

[13] V. M. Muggeo, "Estimating regression models with unknown break-points." *Statistics in Medicine*, vol. 22, pp. 3055–3071, 2003.

[14] L. Yang, S. Liu, S. Tsoka, and L. G. Papageorgiou, "A regression tree approach using mathematical programming," *Expert Systems with Applications*, vol. 78, pp. 347–357, 2017.

[15] C. Dougherty, *Introduction to econometrics*. Oxford University Press, 2011.

[16] P. Vlachos. (2005) StatLib-statistical datasets. *Available at* http://lib.stat.cmu.edu/datasets/.

[17] D. Dheeru and E. Karra Taniskidou. (2017) UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml]. University of California, Irvine, School of Information and Computer Sciences.

[18] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL Data-mining software tool: data set repository, integration of algorithms and experimental analysis framework." *Journal of Multiple-Valued Logic & Soft Computing*, vol. 17, 2011.

[19] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools," *Energy and Buildings*, vol. 49, pp. 560–567, 2012.

[20] I. C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete Research*, vol. 28, pp. 1797 – 1808, 1998.

[21] A. C. Muller and S. Guido, *Introduction to machine learning with python: A guide for data scientists*. O' Reilly Media, Inc., 2016.

[22] GAMS Development Corporation, "General Algebraic Modeling System (GAMS) Release 24.7.1, Washington, DC, USA," 2016.

[23] M. Kuhn, "Building predictive models in R using the caret package," *Journal of Statistical Software*, vol. 28, pp. 1–26, 2008. [Online]. Available: https://www.jstatsoft.org/v028/i05

[24] J. R. Quinlan *et al.*, "Learning with continuous classes," in *5th Australian joint conference on artificial intelligence*, vol. 92. World Scientific, 1992, pp. 343–348.

[25] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," 1996.

[26] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Taylor & Francis, 1984.

[27] Rulequest. (2018) Data Mining with Cubist. *Available at* https://www.rulequest.com/cubist-info.html.

[28] K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer, and A. Zeileis. (2018) Package RWeka. *Available at* https://cran.r-project.org/web/packages/RWeka/RWeka.pdf.

[29] T. Therneau, B. Atkinson, and B. Ripley. (2018) Package rpart. *Available at* https://cran.r-project.org/web/packages/rpart/rpart.pdf.

[30] M. Kuhn, S. Weston, C. Keefer, N. Coutler, and R. Quinlan. (2018) Package Cubist. *Available at* https://cran.r-project.org/web/packages/Cubist/index.html.

[31] B. L. Welch, "The generalization of 'student's' problem when several different population variances are involved," *Biometrika*, vol. 34, pp. 28–35, 1947. [Online]. Available: http://www.jstor.org/stable/2332510

A Spatial Data Analysis Approach for Public Policy Simulation in Thermal Energy Transition Scenarios

Lina Stanzel
AEE - Institut für Nachhaltige Technologien
Gleisdorf, Austria
l.stanzel@aee.at

Johannes Scholz
Graz University of Technology
Graz, Austria
johannes.scholz@tugraz.at

Franz Mauthner
AEE - Institut für Nachhaltige Technologien
Gleisdorf, Austria
f.mauthner@aee.at

Abstract— The paper elaborates on an approach to simulate the effect of public policies regarding thermal energy transition pathways in urban communities. The paper discusses the underlying methodologies of calculating Heating Energy demand of buildings and the rationale for potential zones for thermal energy systems. In order to simulate the effects of public policies on communities the authors developed a spatial Agent-based Model, where the buildings are the main objects that are subject to change, based on a number of both technically and socio-demographic parameters. In order to fill a spatial Agent-based Model with data a number of open source and commercially available datasets need to be spatially analyzed and merged. The initial results of the spatial Agent-based Model simulation show that public policies for thermal energy transition can be simulated accordingly.

Keywords—Agent-based Simulation; Thermal energy transistion; spatial information;

I. INTRODUCTION

International energy and climate policies are aiming at a reduction of global energy demand and carbon dioxide emissions to limit global warming. In response to the Paris Agreement, the European Commission e.g. is aiming for a climate-neutral Europe by 2050 [1], which can only be achieved with a comprehensive transformation of today's energy system. In this respect and driven by the global trend towards urbanization [2], communities and cities are playing an increasingly important role in achieving climate targets and decarbonizing our energy system.

The solution for a low-carbon energy supply of the future, which is widely recognized in science and politics, lies in the combination of an increase in energy efficiency and the use of renewable energy sources [3]. The heating / cooling sector plays a central role here: In Austria, more than half of the total final energy consumption is required to provide space heating, domestic hot water and space cooling (31%) on the one hand and process heat on the other (20%), with fossil fuels being used predominantly (57%) to cover this thermal energy demand [4]. In contrast to the electricity or mobility sector, the provision of thermal energy is predominantly local and decentralized. Hence, for a successful energy transition a successful heat transition at the level of municipalities and cities is essential.

The purpose of this publication is to describe a data analytics methodological approach to generate an agent-based model for public policy simulation of thermal energy transition pathways in urban communities. Special focus is given to the residential building sector which amounts for a significant proportion of both local energy demand as well as local carbon emissions in cities and communities. With the proposed agent-based approach, current urban building energy demand and supply is modelled with spatial reference and future possible transition pathways with respect to building characteristics (e.g. building age, kind of building use, renovation state, heating demand), the availability of local (renewable) energy sources as well as by considering (demographic and empirical) triggers that affect houseowner's decisions towards quality of building renovation and choice of heating systems [5] [6]. The method is being tested for a city in Styria, Austria (Gleisdorf). Following the data analytics part, we elaborate on the spatial Agent-based Modeling and Simulation approach that is developed using the GAMA platform [7].

II. METHODOLOGICAL BACKGROUND & APPROACH

In this paper we elaborate on a methodological approach for an agent-based model and simulation of thermal energy transition pathways in urban communities. In particular we focus on the data analytics and integration necessary to formulate and fuel a spatial agent-based public policy simulation. In the paper we prioritize the residential building sector which amounts for a significant proportion of both local energy demand as well as local carbon emissions in cities and communities. With the proposed agent-based approach, current urban building energy demand and supply is modelled with spatial reference and future possible transition pathways are investigated considering building characteristics (e.g. building age, kind of building use, renovation state, heating demand), the availability of local (renewable) energy sources as well as by considering (demographic and empirical) triggers that affect house owner's decisions towards quality of building renovation and choice of heating systems.

In order to model the effects of public policies in thermal energy transition, we need to consider at least the following information: address points, building polygons, built-up, number of floors, gross floor area (GFA), building use category, building age class, building heating system, building renovation

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_8

state, space heating energy demand (HED), digital elevation model (DEM), land use zones and classes, as well as existing natural gas and district heating networks. Besides these technical information, demographic data play a crucial role in modeling and simulating energy transitions. This is due to the fact that, building renovation or the change of a heating system is –to a large degree – initiated by the owners of a building. Hence, there is a need for demographic information as well [5] [6]. Based on the literature age, education and income level of the owners and/or people residing in a building would be ideal.

From a methodological point of view, that set of information is sufficient to define a spatial Agent-based Model (ABM) and simulate the behavior of agents thereafter. Hence, we describe the components of the ABM first and later elaborate on the spatial data analysis methodology to fill the ABM with data & information.

A. Spatial Agent-based Model

ABMs are a computer-based modeling and simulation approach that use multiple autonomous agents acting in a defined environment [8]–[10]. ABMs can simulate the behavior and dependencies of and between agents, between agents and environment [11]. Similar to natural persons, agents act in a defined environment. Due to the fact that each agent can act independently of other, they may show different behavior ranging from goal-oriented, adaptive to fully autonomous. Generally, ABMs and Agent-based Simulations (ABS) do not reach equilibrium and are not deterministic [8]. Thus, they are capable of answering the question of how a system might react to changing environment. Although agents in ABMs are acting on a micro-level, the results thereof have to be analyzed and visualized on a macro-level – due to the non-deterministic and stochastic nature of agent's decisions [12]. In addition, even if the behavior of agents may be simple in nature, the results may show complex patterns due to the dynamic interactions [13]. ABMs have been widely applied to simulate social, economic and ecological problems [14]–[17].

The ABM and Simulation developed in this research work consists of the agents and an environment. The agents in this ABM are:

- Buildings
    - their attributes
        - Geometry
        - Building type
        - Building period
        - Building age
        - GFA
        - Solar area
        - Fuel category (heating)
        - HED
        - Renovation year
        - Within energy zone
        - Owners
    - Methods
        - Build()
        - Renovate()

The environment is the city of Gleisdorf (Austria), where the buildings are located. In the environment the following entities are to be found:

- Energy infrastructure
    - Attributes
        - Heating type/fuel (e.g. district heating, gas heating)
        - Geometry
    - Methods
        - Grow()
- Energy zones
    - Attributes
        - Type (e.g. gas biomass, heating pump, gas heating, district heating, electric heating, oil, coal)
        - Geometry
    - Methods
        - Update()

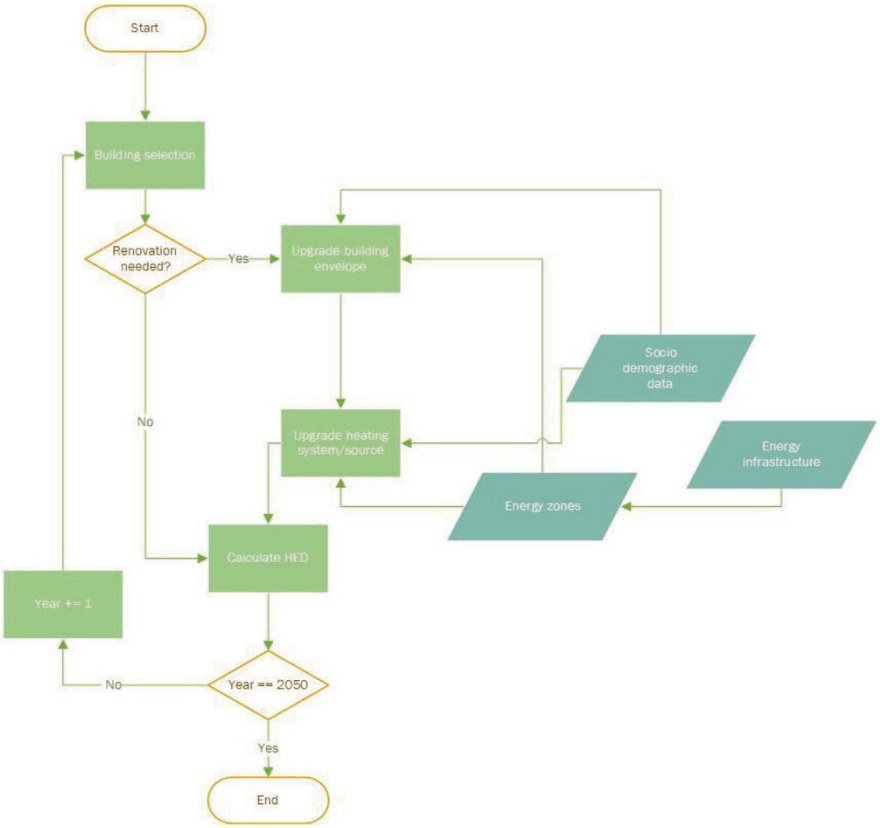
The energy infrastructure and energy zones interact with the buildings. Both environment classes may change over time, as e.g. the energy infrastructure might change their spatial extent (i.e. geometry).

The rationale of the ABM is depicted in Fig. 1. Hence, every year each building is analyzed concerning its renovation status. If the building reaches a certain age it is a candidate for renovation. Renovation of a residential building takes place in two stages, depending on the energy zones and the socio-demographic data that determine the owner's demographic data. In addition, only a certain proportion of candidate buildings are renovated, which is a stochastic process – similar to the real world. If each building has been examined and upgraded (if applicable), the simulation is rolled forward by one year and starts again with the building examination. The simulation stops after the year 2050 to have results at the time of the planned fulfilment of the Paris Climate Agreement.

The resulting parameters from the ABM are as follows:

- Heating system (type & quantity)
- Building renovation information (category of building envelope renovation & renovation year)
- HED/sqm
- HED total

Figure 1: Flow diagram depicting the overall rationale behind the ABM of this paper.



- CO_2 emissions per building
- CO_2 emissions total

Based on these results any public policy (like promoting biomass energy, district heating, etc.) and the thermal energy transition can be assessed and compared – based on the reduction in HED and CO_2 emissions.

B. Modeling of the building heating demand

For the current research area, building thermal energy demand data of the status quo is either derived from available georeferenced cadastral and building energy certificate data or is calculated for defined building archetypes following a simplified one-zone thermal resistance model according to EN ISO 13790 [18]. Future building thermal energy demands are calculated considering different renovation measures for each building archetype according to EN ISO 13790 or, alternatively, by means of dynamic building simulation as introduced by [19]. Building archetypes as well as physical building parameters for this archetypes used for building thermal energy calculation are retrieved from TABULA [20]. Important is the building usage, as residential buildings have a higher HED as industrial or commercial buildings.

C. Modeling of the local energy zones

The energy zones in a city are the areas which have a specific suitability for a certain thermal energy supply system. Within a city there are several geographical areas that are e.g. suitable for thermal heating with gas – or district heating, as the buildings are in the vicinity of the heating or gas network. Other energy zones modeled in this paper describe areas where a certain local heating system is preferable or advisable from both a geographical perspective and/or other parameters, which are described below. In the paper we work with the following sustainable energy zones to force the CO_2 reduction:

- District heating
- Gas heating
- Heating pumps: air source, groundwater source, sole source heating pump
- Solar heating and photovoltaics in combination with heat pumps
- Biomass heating

The modeling of these zones involves spatial information, and spatial relations. The energy zones around existing networks (gas/district heating) is defined to be within a buffer of 35 m (N-Gas) resp. 50 m (district heating) around the networks. An expansion of the district heating network may take place if the heat demand exceeds 35GWh/km^2 which corresponds to a high degree of feasibility according to [21].

The energy zones for heating pumps are defined in a way that they are recommended where energy production with photovoltaic is possible. In addition, the total HED/mGFA^2 should be lower than 30 kWh. Based on the literature the different heating pump technologies are suggested in the following order: sole source, groundwater source, air source [22].

Air source pumps are advisable in areas where noise pollution through the fan of the air source heating pump is not problematic. Hence, the air source heating pump shall not be deployed in areas where a more silent environment is needed (spa areas, graveyard, recreational area). In addition, traffic zones, parking spots and open land are not suited for any heating systems. Groundwater source heating pumps are only allowed in areas with high ground water availability. Sole water source heating pumps can be utilized in areas with a certain level of thermal soil conductivity. Here we use a thermal soil conductivity of greater then 2.05 W/m/K. For thermal solar potential we analyze the roof area in combination with the solar radiation potential according to the solar cadastre Styria. For

small roofs we do not suggest the construction of thermal solar heating.

D. Socio-demographic indicators and empirical triggers considered

Due to the fact that the houseowner's decisions concerning heating energy systems and building renovation is partially based on demographic triggers [6], we include socio-demographic information as well. In particular we have an eye on the education level, the percentage of unemployed persons and the age structure of the people living in the houses. These statistical data was derived from the 250 m demographic raster data of the Statistik Austria. According to the scientific literature the well-educated with higher income level people tend to install ecologic-friendly thermal heating systems and perform a high quality building renovation. In addition, the age structure is relevant for the question "when" a building renovation/heating system replacement takes place. Elder people tend to stick with the contemporary heating system and/or energy source, whereas younger residents are more willing to switch to a "greener" thermal heating system. Similar, younger people are more likely to renovate the building in a high quality manner – thus reducing the HED of the building.

III. Data Sources and Data Analytics for the ABM

In order to generate the information and data for the ABM and the modeling of the thermal energy demand of the buildings, as well as the energy zones a data science and data fusion approach is necessary. This is due to the fact that the necessary information – listed in subsections A, B, C, and D – are not available from stock. Hence, we analyzed a number of open data sets as well as commercially available datasets from relevant public authorities.

We collected the following datasets in order to generate the information needed to "feed" the ABM developed in this paper:

- Cadastre (DKM)
- Address, building and apartment register of Austria (AGWR)
- District heating network data (including buildings) (DH)
- Natural gas network data (GAS)
- Chimney sweeper data (CSD)
- Herold company data (HE)
- ZEUS building energy certification data (ZEUS)
- Energy cadaster (from a local energy provider) (CEP)
- Solar Cadastre Styria
- Socio-Demographic data (DEMO) of 2011 and 2016 from STATISTIK Austria

First, we need to fill the geometry and attributes of the agents of the ABM – i.e. the buildings. Therefore, the topological correctness of the building geometries from the DKM is validated and fixed where necessary. Where no current building data were available, the database was updated to 2018 built-up status according to the Google satellite basemaps with the goal to be able to run spatial analysis methods on the building polygons. To be able to assign addresses to buildings, the address points from the AGWR are matched with the buildings. If an address has more than one building, we merge the single building polygons belonging to one address to a multipolygon. This allows us to access & identify buildings having one address in one single step. Finally, each building needs to be distinguished regarding the usage type – i.e. whether being commercial or residential. As commercial buildings have lower HED we need to identify them accordingly. Therefore the HE data are utilized. By identifying the addresses of companies in HE we are able determine the use type of each building/apartment and store the information accordingly.

For the heating type of each building we use a mix of several data sources. Basically, an overlay operation of buildings with CDP, CSD, DH and GAS is done. Having determined the heating type of each building, the specific HED needs to be calculated, according to subsection B. Nevertheless, for newly renovated (since 2009) and new buildings (since 2008) an energy certification has to be made. The data of the certificates are collected and stored in the ZEUS database, which is used for the purpose of determining the HED of each building. For all other building, the HED needs to be calculated according to subsection B.

For the building type (residential, commercial, etc.) we utilize HE, ZEUS, AGWR, CEP data. From an integration of those datasets based on the spatial component and/or the address, we are able to determine the building type accurately. Additionally, the AGWR data may be used to fill some more attributes of each building such as the usage category, building phase and number of floors. Hence, AGWR data are used to fill attributes of buildings that have not been covered by any other data set so far. Furthermore, we make a cross-check of the building attributes with AGWR data in order to ensure accurate data.

Socio Demographic data are necessary to model the decisions of the agents is obtained from Austrian Census Bureau (Statistik Austria). The data are aggregated socio demographic data, available in raster cells of 250m resolution. The data contain information on the absolute number of citizens, distribution of citizens in age classes, household income, education level (distribution). Due to the fact that the data are available as raster cells we need to disaggregate the data set [23], [24]. The disaggregation process here is based on the building dataset. Hence, we distribute the population present in each 250m raster cell on the available residential buildings and aggregate the innumerable attribute classes to simplify the calculations. Commercial buildings and industrial sites are left out, due to the fact, that people are not "living" in offices or factories (at least not on a regular basis). Furthermore for all of these buildings the values would corrupt the results.

Residents in a raster cell are distributed on the available residential buildings (in fact address points) in the cell with respect to the GFA of a single address. If an address has more than one building we sum up the GFA and allocate residents there. In order to come up with a reliable allocation of inhabitants in terms of social status and age classes, we use a

randomization function to assign age, education level, and employment status (unemployed vs. employed).

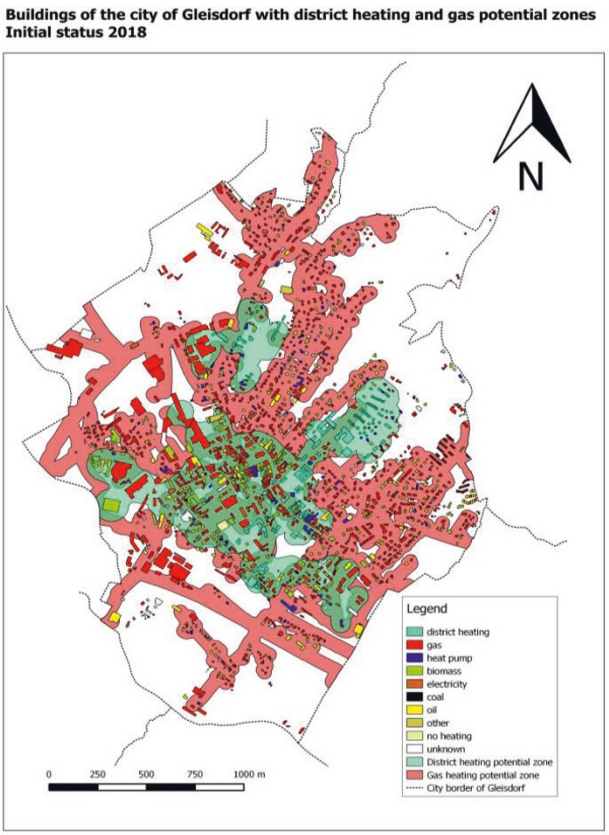


Figure 2: Map of the city of Gleisdorf with the buildings (and their current heating types – as of 2018), district heating network, and potential zones for gas heating and district heating.

The implementation of this data fusion process was done using Quantum GIS[1], PostgreSQL[2] with PostGIS[3], and spatial R[4].

IV. Results

The results of the approach presented here are a spatial ABM that is capable of simulating the effect of public policies regarding thermal energy transition pathways in urban communities. With the data science approach presented here, we are able to generate the basic data necessary for a functioning ABM simulation. The initial results of the ABM are depicted in Fig. 2 and Fig. 3. In Fig. 2 the current situation is given, where the buildings and their heating types are depicted in different colors.

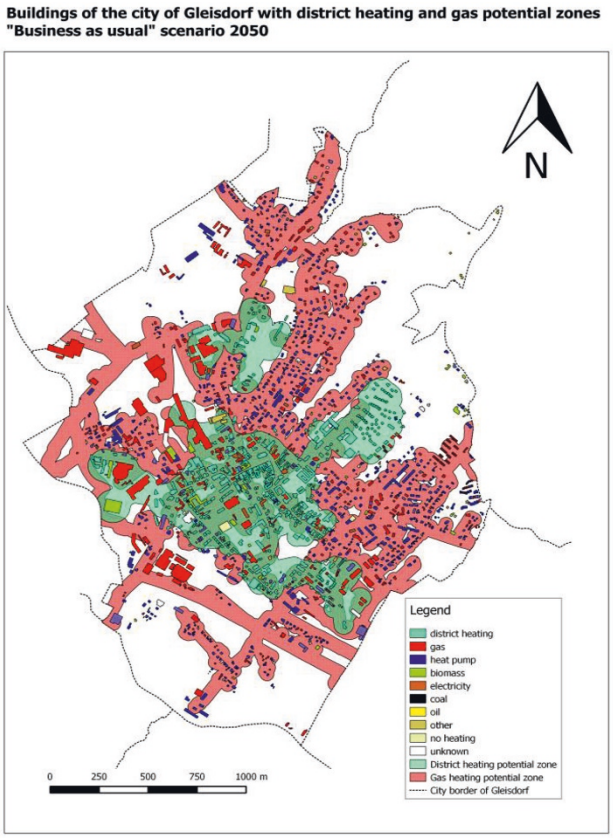


Figure 3: Map of the city of Gleisdorf – the result after the ABM simulation for the year 2050. The maps show the buildings (and their current heating types), district heating network, and potential zones for gas heating and district heating.

Table I.

Buildings per heating type 2018 vs. 2050		
	2018	**2050**
Biomass	110	199
Coal	7	4
District heating	102	192
Electricity	72	19
Gas	897	170
Heat Pump	86	989
Oil	240	36
Other	48	12
No heating	19	4
Unknown	58	14

[1] https://www.qgis.org
[2] https://www.postgresql.org
[3] https://postgis.net
[4] https://www.rspatial.org

In total we are looking at 1639 buildings in the current situation and in 2050. After the spatial ABM simulation – where we assume that politicians follow an ambitious eco-friendly agenda – several homeowners have switched to an environmental friendly heating system. The detailed results of heating type update for the usual scenario setup are depicted in Tab. I. For this scenario the average HED per qm was reduced by 38 percent from 123 to 74 kWh and the CO_2 Equivalent in kg CO_2/kWh was reduced from 23 to 9 kg. This is a reduction of estimated 62 per cent. The results reveal that the spatial ABM simulation is capable of simulating the political agenda - nevertheless following guidelines and rules for thermal energy heating. Hence, the city cannot force the usage a certain heating type or the renovation of the peoples building envelopes, but can make a "greener" energy usage more interesting – e.g. by incentives.

V. Summary & Discussion

The paper elaborates on an approach to simulate the effect of public policies regarding thermal energy transition pathways in urban communities. The authors presented the underlying methodologies of calculating HED of buildings and the rationale for potential zones for thermal energy systems. In order to simulate the effects of public policies on communities the authors developed a spatial ABM, where the buildings are the main objects that are subject to change, based on a number of both technically and socio-demographic parameters. Hence, homeowners may decide on a certain building renovation, based on the environment and the socio-demographic context of themselves. In order to fill a spatial ABM with data a number of open source and commercially available datasets need to be spatially analyzed and merged. The initial results of the spatial ABM simulation shows that public policies for thermal energy transition can be simulated accordingly.

In the future, different possible scenarios will be developed on the basis of hypotheses which are based on socio-demographic research of [5], [6].

These scenarios are intended to cover different policy decisions. Furthermore, these different developments should be made comparable with the initial status and with each other. In the future the simulation will also be applied to the surrounding communities of the city of Gleisdorf which, have a more rural character.

References

[1] European Commission, "A Clean Planet for all - A European strategic long-term vision for a prosperous, modern, competitive and climate neutral economy," COM(2018) 773, 2018.

[2] United Nations, "World Urbanization Prospects 2018," United Nations, Department of Economic and Social Affairs, Population Division, 2018.

[3] H. Lund, "Renewable energy strategies for sustainable development," *Energy*, vol. 32, no. 6, pp. 912–919, Jun. 2007.

[4] G. Günsberg, A. Veigl, and J. Fucik, "Faktencheck Energiewende 2018 / 2019," Klima- und Energiefonds Austria, Vienna, 2018.

[5] T. Decker and K. Menrad, "House owners' perceptions and factors influencing their choice of specific heating systems in Germany," *Energy Policy*, vol. 85, pp. 150–161, Oct. 2015.

[6] M. Hecher, S. Hatzl, C. Knoeri, and A. Posch, "The trigger matters: The decision-making process for heating systems in the residential building sector," *Energy Policy*, vol. 102, pp. 288–306, Mar. 2017.

[7] E. Amouroux, T.-Q. Chu, A. Boucher, and A. Drogoul, "GAMA: An Environment for Implementing and Running Spatially Explicit Multi-agent Simulations," in *Agent Computing and Multi-Agent Systems*, vol. 5044, A. Ghose, G. Governatori, and R. Sadananda, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 359–371.

[8] A. T. Crooks and A. J. Heppenstall, "Introduction to agent-based modelling," in *Agent-based models of geographical systems*, A. Heppenstall, A. T. Crooks, L. M. See, and M. Batty, Eds. Springer, 2012, pp. 85–105.

[9] P. Mandl, "Multi-Agenten-Simulation und Raum - Spielwiese oder tragfähiger Modellierungsansatz in der Geographie," *Klagenfurter Geogr. Schriften*, no. 23, pp. 5–34, 2003.

[10] M. Batty, A. Crooks, L. See, and A. Heppenstall, "Perspectives on Agent-Based Models and Geographical Systems," pp. 1–15, 2012.

[11] B. Edmonds, "The use of models-making MABS more informative," in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, 2000, pp. 15–32.

[12] C. Macal and M. North, "Tutorial on agent-based modelling and simulation," *J. Simul.*, vol. 4, no. 3, pp. 151–162, 2010.

[13] N. Gilbert and K. Troitzsch, *Simulation for the social scientist*. McGraw-Hill Education (UK), 2005.

[14] N. M. Gotts, J. G. Polhill, and A. N. R. Law, "Agent-based simulation in the study of social dilemmas," *Artif. Intell. Rev.*, vol. 19, no. 1, pp. 3–92, 2003.

[15] S. Heckbert, T. Baynes, and A. Reeson, "Agent-based modeling in ecological economics," *Ann. N. Y. Acad. Sci.*, vol. 1185, no. 1, pp. 39–53, Jan. 2010.

[16] M. B. Marietto, N. David, J. S. Sichman, and H. Coelho, "A classification of paradigmatic models for agent-based social simulation," in *International Workshop on Multi-Agent Systems and Agent-Based Simulation*, 2003, pp. 193–208.

[17] P. Davidsson, "Agent based social simulation: A computer science view," *J. Artif. Soc. Soc. Simul.*, vol. 5, no. 1, 2002.

[18] ISO, "13790: 2008 Energy performance of buildings–Calculation of energy use for space heating and cooling," *Int. Stand. Organ.*, 2008.

[19] P. Nageler *et al.*, "Novel validated method for GIS based automated dynamic urban building energy simulations," *Energy*, vol. 139, pp. 142–154, Nov. 2017.

[20] T. Loga, B. Stein, and N. Diefenbach, "TABULA building typologies in 20 European countries—Making energy-related features of residential building stocks comparable," *Energy Build.*, vol. 132, pp. 4–12, Nov. 2016.

[21] B. Möller, E. Wiechers, U. Persson, L. Grundahl, and D. Connolly, "Heat Roadmap Europe: Identifying local heat demand and supply areas with a European thermal atlas," *Energy*, vol. 158, pp. 281–292, 2018.

[22] V. Quaschning, *Regenerative Energiesysteme: Technologie–Berechnung–Simulation, 7. aktualisierte Auflage*. Muenchen: Hanser FAchbuchverlag, 2011.

[23] F. Gallego, F. Batista, C. Rocha, and S. Mubareka, "Disaggregating population density of the European Union with CORINE land cover," *Int. J. Geogr. Inf. Sci.*, vol. 25, no. 12, pp. 2051–2069, 2011.

[24] J. Scholz, M. Andorfer, and M. Mittlboeck, "Spatial Accuracy Evaluation of Population Density Grid Disaggregations with Corine Landcover," in *Geographic Information Science at the Heart of Europe*, Springer, 2013, pp. 267–283.

Data Analytics | Comprehensibility

A Probabilistic Approach to Web Waterfall Charts

Maciej Skorski
DELL Slovakia
Bratislava, Slovakia
maciej.skorski@gmail.com

***Abstract*—The purpose of this paper is to propose an *efficient and rigorous* modeling approach for *probabilistic waterfall charts* illustrating timings of web resources, with particular focus on fitting them on *big data*. An implementation on real-world data is discussed, and illustrated on examples.**

The technique is based on non-parametric density estimation, and we discuss some subtle aspects of it, such as noisy inputs or singular data. We also investigate optimization techniques for numerical integration that arises as a part of modeling.

***Index Terms*—probabilistic waterfall charts, non-parametric density estimation**

I. Introduction

A. Waterfall charts

Waterfall charts are a fundamental tool to diagnose web pages. They allow developers to understand the execution flow, identify bottlenecks and suggest possible improvements. Basically, waterfall charts illustrate start and end times of *resources* (html, css, java scripts, api calls and others) that a web page loads; as they don't run fully sequentially but rather in asynchronous batches, the chart looks like a cascade (see examples at the end of the paper) hence the name.

The vast majority of online literature discusses the charts in a *deterministic setting* (one page load), but in reality the chart is going to be different with each attempt (because of network latency, cache etc) and vary even more across page visitors (due to different devices, operation systems, web browsers, connection quality). This variability can be partially reduced by restricting the analysis to a specific group of loads (e.g. same device and browser), yet there are still many factors not measured and reported, which makes timings fluctuate.

To analyze this variability and get more accurate insights into ''average'' customer experience, one would need *probabilistic waterfall charts* which for each page resource illustrate variation of start end ending times. Particular challenges that comes to mind are

- How to *actually* visualize a three-dimensional cascade of timings (resource, start, end)?
- What model to use to maintain flexibility and possibly low bias (volume of traffic, especially after restrictions applied, may not be big enough to model all resources with high confidence)?
- How to make the model possibly *agnostic of domain knowledge* (for example *not* relying on any explicit dependency structure on resources)?
- How to fit the model on *raw big data*, e.g. on a repository of data captured by Yahoo's Boomerang software [1]

In the remainder of the paper we will discuss a model that attempts to address these issues.

The full version available on arxiv

B. Related works

The Critical Path Method [2] is widely used in project management to understand the shortest possible duration of the project which depends on events with dependencies (e.g. item C waits until A and B are done). In our setting this approach would be basically about finding *a sequential chain of resources* with high total duration. The problem with this approach is that it relies on the knowledge of dependencies, and that we want also to illustrate overlapping timings.

It is hard to find literature which discusses probabilistic waterfall charts together with mathematical foundations. We believe that a study case discussed together with mathematical modeling foundations will be of benefit for data scientists that will come across this problem.

C. Organization

The paper is organized as follows: Section II explains the model and discusses its theoretical properties. Section III discusses the implementation and Section IV presents example outcomes from the real-world implementation. Finally Section V concludes the work.

II. Model

A. Starts and Ends in One: Execution Probability

For a fixed resource r, consider a joint distribution of pairs $(x, y) = (t_{\text{start}}, t_{\text{duration}})$. When modeling we prefer to work with duration rather than end time, because it is more independent of start times. At the heat of our modeling approach is the following question

what are the odds that a resource r runs at time t?

then answer to which we quantify by the metric called the *execution probability*. It is defined as

$$p_{\text{exec}}(r;t) = \int\limits_{\substack{0\leqslant y \\ x\leqslant t\leqslant x+y}} p_r(x,y)\mathrm{d}(x,y) \tag{1}$$

The definition is visualized in Figure 1.

With the metric defined in Equation (1) one can retrieve the range of variation. As a function of t it assumes typically a beta-like shape as illustrated in Figure 2 Simple inference is

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_9

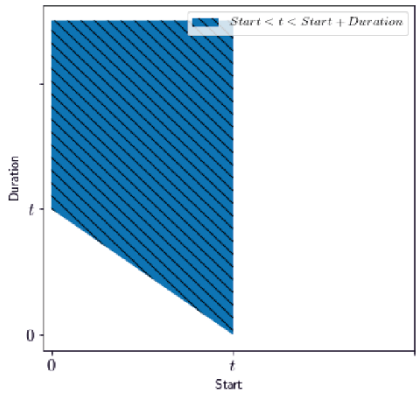


Fig. 1. Execution probability, definition. A given time point t must be between the start and end times, which corresponds to a trapezoid area in terms of the (start, duration) variables.

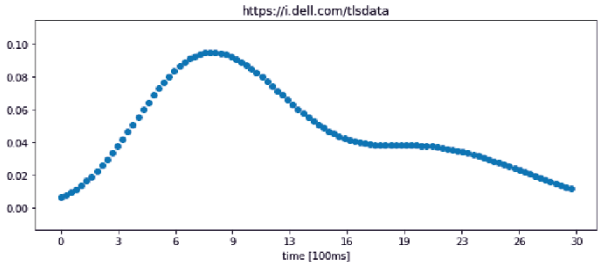


Fig. 2. Execution probability as a function of time (two-dimensional distribution of pairs (start, duration) has been estimated by KDE).

also possible for comparison of pairs of resources (although we don't model the dependency structure!), by looking at the overlap of execution probabilities: when it is small, the resources run (almost) sequentially. Simple credibility or confidence bounds can be derived, but we do not discuss that here in the paper.

Note that (1) can be theoretically defined for any distribution, also for singular distributions (e.g. when one resource has no variation, we will discuss this issue later). The proposed construction of waterfall charts is to visualize $p_{\text{exec}}(r;t)$ in two dimensions: vertically across resources r and horizontally on a grid of points t. This visualizes triples $(r, t_{\text{start}}, t_{\text{duration}})$ on two dimensions (see examples in Section IV).

B. Density Estimation

Having established the methodology to visualize distributions, we move to the problem of density estimation. In order to compute $p_{\text{exec}}(r;t)$ defined in Equation (1) we have to model the joint distribution of start and end times $p_r(x,y)$. Parametric families like beta distributions are natural candidates, but are not flexible enough. Indeed, in our experiments we have observed resources with multi-modal shapes. The natural alternative is non-parametric modeling by *kernel density estimation* developed by Rosenblatt and Parzen [3], [4]. It is defined as the *convolution* of the empirical distribution of data $\mathcal{D}$ and a kernel function K (both d-dimensional)

$$p(\mathbf{x}) = \left(\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \delta_{\mathbf{x}_i} \right) \star K_{\mathbf{H}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_i \in \mathcal{D}} \int K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}') \delta_{\mathbf{x}_i}(\mathbf{x}') \mathrm{d}\mathbf{x}' \quad (2)$$

where $\delta_{\mathbf{x}_i}$ is the unit mass distribution at $\mathbf{x}_i$, $\mathbf{H}$ is the matrix parameter called *bandwidth* or *scale* (selected by optimization rules discussed later) and

$$K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{x})$$

is the scaled version of the base kernel K; here $|\mathbf{H}|$ denotes the absolute determinant of $\mathbf{H}$ and $\mathbf{H}^{-1}$ is the matrix inverse. Although the bandwidth $\mathbf{H}$ can be pretty arbitrary, it is a common practice to make it *similar to data* by scaling the data covariance $\mathbf{H} = h \cdot \mathrm{cov}(\mathcal{D})$ with a positve parameter h.

The kernel function K must be itself a symmetric probability distribution with sufficiently fast decay; the popular choice is the Gaussian kernel. Multidimensional kernels are also often build as the product $K(\mathbf{u}) = \prod_i K^0(\mathbf{u}_i)$ out of a single-dimensional kernel K^0, which we refer to as product kernels. Under some smoothness assumptions on the true density, with growing data size and the bandwidth selected appropriately (cross-validation on a dataset or asymptotically formulas are usually used), the kde converges to the true data distribution with much lower bias rate than histograms. We refer to [5] for an excellent survey.

In our case we have a dataset $\mathcal{D}$ which consists of points $(r, t_{\text{start}}, t_{\text{duration}})$, and for each fixed resource r consider the 2-dimensional kde estimate of the set of the resource timings $\mathcal{D}_r = \{t_{\text{start}}, t_{\text{duration}} : (r, t_{\text{start}}, t_{\text{duration}}) \in \mathcal{D}\}$. We thus define

$$p_r(x,y) = \left(\frac{1}{|\mathcal{D}_r|} \sum_{(x_i,y_i) \in \mathcal{D}_r} \delta_{x_i,y_i} \right) \star K_{\mathbf{H}} = \frac{1}{|\mathcal{D}_r|} \sum_{(x_i,y_i) \in \mathcal{D}_r} \int K_{\mathbf{H}}(x-x', y-y') \delta_{x_i,y_i}(x',y') \mathrm{d}(x',y') \quad (3)$$

As the base kernel we use *the gaussian density*, defined as

$$K(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp(-\mathbf{x}^T\mathbf{x}/2)$$

for d-dimensional $\mathbf{x}$ ($d = 2$ in our case).

C. Data Aggregation

In order to fit the density estimation we need a sample of data points. However, *we do not subsample* (taking a sample representative for all resources may be a bit subtle and actually unnecessary complicated). Instead, we *round and aggregate the data* (on Hive) as shown in Listing 1 (the code

is simplified as in reality we use more filtering conditions, for example to distinguish between cached and hard page loads). This gives us frequencies of rounded timings for each resource r (we round up to integer multiplicities of 100ms). As the data has been discretized and compressed by rounding, we can afford efficient modeling. What is important in this simple technique is itself a *simple form of KDE*.

Proposition 1 (Data aggregation as preliminary KDE)**.** *The process in Listing 1 is for each resource a kde estimate with rectangular kernel and bandwidth of* 100*.*

The proof appears in Appendix A. As a consequence we see that the aggregation processes copes well with further kde modeling (a composition of two kde estimations is again a kde estimator, by general properties of the convolution operation). Essentially the effect is as if we were working on the full data set, with only small (acceptable in our applications) bias due to rounding.

Listing 1. Data Aggregation on Hive

```
-- timings stored in miliseconds
select
  resource,
  round(start_time/100) start,
  round((end_time-start_time)/100)) duration,
  count(1) hits
from timings_data
group by
  resource,
  round(start_time/100),
  round((end_time-start_time)/100))
```

D. Handling Singularities

In some cases data points are linearly dependent and the kde estimator Equation (3) is not well defined because the covariance matrix is singular. A typical case is a resource with very small duration, which is captured (during measurment or later rounded by post-processing) as 0. To handle such scenarios we add a *small ammount of noise* to the data and slightly modify Equation (1). Namely, we redefine

$$\mathcal{D} = \mathcal{D} + \mathbf{n} \tag{4}$$

where $\mathbf{n}$ is a vector of uniform samples from $[-\epsilon, \epsilon]$ of same shape as $\mathcal{D}$, and change Equation (1) to

$$p_{\text{exec}}(r;t) = \int_{-\epsilon}^{t+\epsilon} \int_{t-x-\epsilon}^{\infty} p_r(x,y)\mathrm{d}y\mathrm{d}x \tag{5}$$

(this is essentially extending boundaries of the area in Figure 2 by $O(\epsilon)$). The parameter ϵ needs to be sufficiently small, adjusted depending on the scale $\mathbf{H}$ as discussed below.

We will prove that this approach produces a good approximation. First we observe that for sufficiently small ϵ the difference between Equation (5) and Equation (1) under the kde estimator is small.

Proposition 2 (Modified integration of kde)**.** *If $p_r(x,y)$ is a 2-dimensional kde estimate, then the difference between Equation (5) and Equation (1) is $O(p_r^{\max}(x) \cdot \epsilon)$ where $p_r^{\max}(x)$ is the mode of the marginal distribution $p_r(x)$. For the gaussian case, we have $O(\epsilon/\|\mathbf{H}\mathbf{H}^T\mathbf{e}_1\|)$ (the norm being the Euclidean norm).*

The proof of Proposition 2 appears in Appendix C.

For the two-dimensional case it remains to prove that the kde estimator produces close results when noise is added. We state and prove it under general (weak) assumptions as it rigorously justifies the popular heuristic and is interesting on its own.

Proposition 3 (KDE is robust under small input noise)**.** *Suppose that the kernel K has absolutely integrable first derivatives (true for sufficiently fast decay, particularly for gaussian kernels). Then adding noise uniform on $[-\epsilon, \epsilon]^2$ to the data $\mathcal{D}$ changes the distribution in Equation (3) at most by $O\left(\left\|\mathbf{H}^{-1}\right\| \cdot \epsilon\right)$ in total variation.*

The proof is given in Appendix B. Fro two distributions δ-close in total variations probabilities of all sets differ at most by δ. Thus, for sufficiently small ϵ, kde estimates computed on original and respectively noisy inputs give similar integration results.

Finally the singular case is handled by the following fact

Proposition 4 (Model works in one dimension)**.** *If the data in (4) is of the form $\mathcal{D} = (\mathbf{X}, \mathbf{n})$ (respectively $\mathcal{D} = (\mathbf{n}, \mathbf{Y})$) and K is a product kernel then $p_{\text{exec}}(r;t)$ in Equation (5) is equal to a 1-D kde estimate of $\mathbf{Y}$ (respectively: of $\mathbf{X}$) with same base kernel and the scale being the first (respectively: second) diagonal element of $\mathbf{H}$.*

The proof appears in Appendix D. If the original data is of shape $\mathcal{D} = (\mathbf{X}, \mathbf{0})$, after Equation (4) it becomes $(\mathbf{X}+\mathbf{n}_1, \mathbf{n}_2)$ where $\mathbf{n} = (\mathbf{n}_1, \mathbf{n}_2)$ is the noise vector. Then Proposition 4 shows that (5) equals a kde estimate on the 1-dimensional data $\mathbf{X} + \mathbf{n}_1$ and then Proposition 3 shows that this approximates a kde on the denoised data $\mathbf{X}$.

E. Resource Impact on Time-to-Interactive

The Time to Interactive metric (TTI) measures when a page becomes interactive.

NadarayaWatson

III. Implementation Details

We have built the prototype using the following Python packages

- `pandas` [6] to mange data sets (filter,grouping)
- `numpy` [7] and `scipy` [8] for modeling (the kde algorithm comes from `scipy.stats.gaussian_kde`)
- `multiprocessing` [9] to parallelize computations over different resources

The data is captured by Boomerang on company's web pages and is stored on Hive.

A. Speeding up KDE

The kde estimator, as defined in Equation (2), can be simply fit by passing a list of all points in the data set, including repetitions. However this would significantly impact computations and we wouldn't benefit of the aggregation done

Listing 2. Execution probability by Monte-Carlo

```
def p_exec_mc(k,t):
  # k is a two-dimensional KDE object and t is an array
  N = 1000000
  sample = k.resample(N).T
  grid=np.repeat(np.expand_dims(t,0),N,0)
  return ((np.expand_dims(sample[:,0],1) < grid) & \\
  (grid < np.expand_dims(sample.sum(1),1))).sum(0)/N
```

in Listing 1. In fact to fit the estimator it suffices to know *distinct points with frequencies*, and this feature is available since the version 1.2 of Scipy (newest).

B. Execution Probability Computation

The integral in Equation (1) (and similarly in Equation (5)) needs to be evaluated across three dimensions: (a) on a *grid of time points* t (b) for each resource and its corresponding density $p_r(x,y)$ and (c) for each page within the scope of analysis.

To reduce the running time, we have considered the following techniques

- *transform the integral*: we can rewrite Equation (1) as

$$p_{\text{exec}}(r;t) = \int_0^t \int_0^{\infty} p_r(x,y)\mathrm{d}y\mathrm{d}x - \int_0^t \int_0^{t-y} p_r(x,y)\mathrm{d}y\mathrm{d}x$$

  This represents the region of integration in Figure 1 as a difference of the rectangle and triangle. The gain is that the implementation of `scipy.stats.gaussian_kde` natively *supports fast integration over rectangles*; thus the integration subroutine is called only over the triangle area, which is small for many points t.
- *Monte Carlo*: rather than numerically integrating over the area in Figure 1, we can sample from the density and count points hitting the area in Figure 1. This can be implemented efficiently with one large sample broadcast to the entire grid, see Listing 2.
- *parallel computation*: to further increase the processing speed, we process different resources in parallel using the multiprocessing Python library.

IV. Application Examples

We discuss two examples from real data (more will be presented in the full version of the paper)

The first example is illustrated in Figure 3. The resource loaded first becomes a bottleneck, this suggests to start next calls asynchronously.

In the next example, illustrated in Figure 4, we were able to find three external scripts (nexus calls, in yellow) being bottlenecks of the page load.

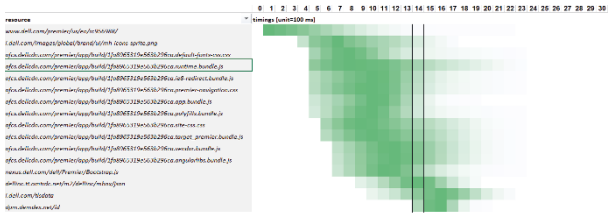


Fig. 3. Example of a probabilistic waterfall chart

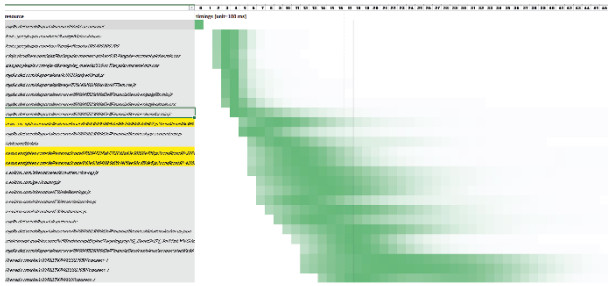


Fig. 4. Example of a probabilistic waterfall chart

V. Conclusion

We have established mathematical foundations of modeling probabilistic waterfall charts for web page resources. An efficient implementation in Python has been discussed and illustrated with real-world examples.

References

[1] Yahoo, "Boomerang," 2011, [Online; accessed 2019-01-24]. [Online]. Available: https://github.com/yahoo/boomerang

[2] J. E. Kelley, Jr and M. R. Walker, "Critical-path planning and scheduling," in *Papers Presented at the December 1-3, 1959, Eastern Joint IRE-AIEE-ACM Computer Conference*, ser. IRE-AIEE-ACM '59 (Eastern). New York, NY, USA: ACM, 1959, pp. 160–173. [Online]. Available: http://doi.acm.org/10.1145/1460299.1460318

[3] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *The Annals of Mathematical Statistics*, vol. 27, no. 3, pp. 832–837, Sep. 1956. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177728190

[4] E. Parzen, "On estimation of a probability density function and mode," *Ann. Math. Statist.*, vol. 33, no. 3, pp. 1065–1076, 09 1962. [Online]. Available: https://doi.org/10.1214/aoms/1177704472

[5] D. W. Scott and S. R. Sain, "Multi-dimensional density estimation," http://www.stat.rice.edu/ scottdw/ss.nh.pdf, 2004.

[6] [Online; accessed 2019-01-24]. [Online]. Available: https://pandas.pydata.org

[7] T. Oliphant, "NumPy: A guide to NumPy," USA: Trelgol Publishing, 2006–, [Online; accessed 2019-01-24]. [Online]. Available: http://www.numpy.org/

[8] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed 2019-01-24]. [Online]. Available: http://www.scipy.org/

[9] "Multiprocessing-Python package," [Online; accessed 2019-01-24]. [Online]. Available: https://docs.python.org/3.4/library/multiprocessing.html

APPENDIX

A. Proof of Proposition 1

The rectangular 1-dimensional kernel is given by $K^0(u) = 1$ when $|u| \leqslant \frac{1}{2}$ and 0 otherwise. In two dimensions we put $K(x,y) = K^0(x) \cdot K^0(y)$. Let $\mathbf{H} = \frac{1}{100} \cdot I$ where I is the 2-dimensional identity matrix. When evaluating the convolution of the data set empirical distribution and $K_{\mathbf{H}}$ on a point $\mathbf{x}$, we simply compute the (normalized) number of data points $\mathbf{x}'$ that satisfy $|\mathbf{x} - \mathbf{x}'| \leqslant \frac{100}{2}$. But this is exactly (up to the normalizing constant) the functionality of the code in Listing 1.

B. Proof of Proposition 3

Proof. Recall that the total variation of densities f and g is defined by $d_{\mathrm{TV}}(f,g) = \int |f - g|$. Observe first that

$$\begin{aligned}&\delta_{x_i,y_i} \star K_{\mathbf{H}}(x,y) - \delta_{x_i+n_i^x,y_i+n_i^y} \star K_{\mathbf{H}}(x,y)\\ &= K_{\mathbf{H}}(x - x_i, y - y_i) - K_{\mathbf{H}}(x - (x_i + n_i^x), y - (y_i + n_i^y))\end{aligned}$$

By the Integral Mean Value Theorem

$$\begin{aligned}&\delta_{x_i,y_i} \star K_{\mathbf{H}}(x,y) - \delta_{x_i+n_i^x,y_i+n_i^y} \star K_{\mathbf{H}}(x,y)\\ &\qquad = \int_0^1 \partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] \mathrm{d}\zeta \qquad (6)\end{aligned}$$

We will prove that the right hand side of Equation (6) is integrable as the function of (x,y), with the upper bound $O(\|\mathbf{H}^{-1}\| \cdot \epsilon)$. For every measurable compact set E we have

$$\begin{aligned}&\int_E \int_0^1 \partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] \mathrm{d}\zeta \mathrm{d}(x,y) =\\ &\qquad \int_0^1 \int_E \partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] \mathrm{d}(x,y) \mathrm{d}\zeta\end{aligned}$$

Let $(x', y') = (x + \zeta n_i^x, y + \zeta n_i^y)$. We have

$$\partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] = DK_{\mathbf{H}}(x', y') \cdot (n_i^x, n_i^y)^T.$$

where D is the first derivative matrix. Thus

$$\begin{aligned}&\int_E \int_0^1 \partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] \mathrm{d}\zeta \mathrm{d}(x,y) =\\ &\qquad \int_0^1 \int_E DK_{\mathbf{H}}(x', y') \cdot (n_i^x, n_i^y)^T \mathrm{d}(x,y) \mathrm{d}\zeta\end{aligned}$$

We can transform the integration changing the variables from (x,y) to (x', y'). Since ζ, n_i^x, n_i^y don't depend on (x,y)

$$\begin{aligned}&\int_E \int_0^1 \partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] \mathrm{d}\zeta \mathrm{d}(x,y) =\\ &\qquad \int_0^1 \int_{E + \zeta(n_i^x, n_i^y)} DK_{\mathbf{H}}(x,y) \cdot (n_i^x, n_i^y)^T \mathrm{d}(x,y) \mathrm{d}\zeta\end{aligned}$$

Which is upper bounded by

$$\begin{aligned}&\left| \int_E \int_0^1 \partial_\zeta \left[K_{\mathbf{H}}(x + \zeta n_i^x, y + \zeta n_i^y)\right] \mathrm{d}\zeta \mathrm{d}(x,y) \right| \leqslant\\ &\qquad \epsilon \int_{E + \epsilon[-1,1]^2} \|DK_{\mathbf{H}}(x,y)\| \,\mathrm{d}(x,y)\end{aligned}$$

Since $K_{\mathbf{H}}(x,y) = \frac{1}{|\mathbf{H}|} K(\mathbf{H}^{-1}(x,y)^T)$ we have

$$DK_{\mathbf{H}}(x,y) = \frac{1}{|\mathbf{H}|} DK(\mathbf{H}^{-1}(x,y)^T)\mathbf{H}^{-1}$$

applying the change of variables $(x,y) := \mathbf{H}^{-1}(x,y)^T$ and extending the integration to the entire space we obtain

$$\begin{aligned}\|DK_{\mathbf{H}}(x,y)\| &\leqslant \frac{1}{|\mathbf{H}|} \left\|DK(\mathbf{H}^{-1}(x,y)^T\right| \left\|\mathbf{H}^{-1}\right\|\\ &\leqslant \epsilon \cdot \|\mathbf{H}^{-1}\| \cdot \int \|DK(x,y)\| \,\mathrm{d}(x,y)\end{aligned}$$

which is $O(\|\mathbf{H}^{-1}\| \cdot \epsilon)$ as DK is assumed to be integrable.

Once we have obtained the bound

$$\begin{aligned}&\int \left|\delta_{x_i,y_i} \star K_{\mathbf{H}}(x,y) - \delta_{x_i+n_i^x,y_i+n_i^y} \star K_{\mathbf{H}}(x,y)\right| \mathrm{d}(x,y)\\ &\qquad \leqslant O(\|\mathbf{H}^{-1}\| \cdot \epsilon)\end{aligned}$$

which gives the total variation between distributions $\delta_{x_i,y_i} \star K_{\mathbf{H}}(x,y)$ and $\delta_{x_i+n_i^x,y_i+n_i^y} \star K_{\mathbf{H}}(x,y)$, the bound on the total variation of $\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i \in \mathcal{D}} \delta_{x_i,y_i} \star K_{\mathbf{H}}(x,y)$ and $\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i \in \mathcal{D}} \delta_{x_i+n_i^x,y_i+n_i^y} \star K_{\mathbf{H}}(x,y)$ follows by the triangle inequality.

C. Proof of Proposition 2

Proof. The difference is at most $2p_{\max}\epsilon$ where $p_{\max}$ is the mode of $p_r(x,y)$; by Equation (3) we have $p_{\max} \leqslant \sup K_{\mathbf{H}} = O(|\mathbf{H}|^{-1})$. Summing up, the difference is bounded by $O(|\mathbf{H}|^{-1}\epsilon)$, the constant depends on the base kernel K. We give a proof in two steps. We first quantify the approximation

$$\int_{-\epsilon}^{t+\epsilon} \int_{t-x-\epsilon}^{\infty} p_r(x,y) \mathrm{d}y \mathrm{d}x \approx \int_{-\epsilon}^{t+\epsilon} \int_{t-x}^{\infty} p_r(x,y) \mathrm{d}y \mathrm{d}x$$

with respect to ϵ. Note that

$$\begin{aligned}&\left| \int_{t-x-\epsilon}^{\infty} p_r(x,y) \mathrm{d}y - \int_{t-x}^{\infty} p_r(x,y) \mathrm{d}y \right|\\ &\qquad = \left| \int_0^{+\epsilon} p_r(x, t - x - \zeta) \mathrm{d}\zeta \right|\end{aligned}$$

integrating over x gives

$$\begin{aligned}&\int \left| \int_{t-x-\epsilon}^{\infty} p_r(x,y) \mathrm{d}y - \int_{t-x}^{\infty} p_r(x,y) \mathrm{d}y \right| \mathrm{d}x\\ &\qquad = \int \left| \int_0^{+\epsilon} p_r(x, t - x - \zeta) \mathrm{d}\zeta \right| \mathrm{d}x\\ &\qquad \leqslant \Pr_{x,y \sim p_r} [t - 2\epsilon \leqslant x \leqslant t + 2\epsilon]\end{aligned}$$

The error in our approximation is thus $O(p_r^{\max}(x) \cdot \epsilon)$, where $p_r^{\max}(x)$ is the mode of marginal x. Next we consider the approximation

$$\int_{-\epsilon}^{t+\epsilon} \int_{t-x}^{\infty} p_r(x,y) \mathrm{d}y \mathrm{d}x \approx \int_0^{t} \int_{t-x}^{\infty} p_r(x,y) \mathrm{d}y \mathrm{d}x$$

The error is

$$\left|\int_{-\epsilon}^{t+\epsilon}\int_{t-x}^{\infty} p_r(x,y)\mathrm{d}y\mathrm{d}x - \int_{0}^{t}\int_{t-x}^{\infty} p_r(x,y)\mathrm{d}y\mathrm{d}x\right| \leqslant \Pr_{x,y\sim p_r}[x\in[-\epsilon,0]\cup[t,t+\epsilon]]$$

again bounded by $O(p_r^{\max}(x)\cdot\epsilon)$.

We will give a more explicit expression on the error $O(p_r^{\max}(x)\cdot\epsilon)$. To this end we consider the marginal density

$$\int \delta_{x_i,y_i}\star K_{\mathbf{H}}(x,y)\mathrm{d}y = \int K_{\mathbf{H}}(x-x_i,y-y_i)\mathrm{d}y \quad (7)$$

Let $\mathbf{x}_i'=(x-x_i,y-y_i)$. For the gaussian case we have

$$K_{\mathbf{H}}(x-x_i,y-y_i) = \frac{1}{|\mathbf{H}|}\exp\left(-\frac{1}{2}\mathbf{x}_i'^{T}\left(\mathbf{H}^{-1^T}\mathbf{H}^{-1}\right)\mathbf{x}_i'\right)$$

This is the gaussian distribution with covariance matrix $\mathbf{\Sigma}=\mathbf{H}^T\mathbf{H}$, and by integrating over y - equivalently over $\mathbf{x}_i'^2=y-y_i$ - we marginalize its second component. As a result we are left with the first component, which is gaussian with variance $(\mathbf{\Sigma}\mathbf{e}_1)^2$ and thus has mode $\|\mathbf{\Sigma}\mathbf{e}_1\|$. This bounds the mode of $\int\delta_{x_i,y_i}\star K_{\mathbf{H}}(x,y)\mathrm{d}y$ independently on i, hence also the mode of $\int\frac{1}{|\mathcal{D}_r|}\sum_{(x_i,y_i)\in\mathcal{D}_r}\delta_{x_i,y_i}\star K_{\mathbf{H}}(x,y)\,\mathrm{d}y$ which is the marginal distribution $p_r(x)$ of the kde estimate $p_r(x,y)$.

D. Proof of Proposition 4

Proof. Consider the case when the dimension $\mathbf{X}$ is non-zero (the other case is analogues).

$$\mathrm{cov}(\mathbf{X},\mathbf{n}) = \begin{pmatrix}\mathrm{Var}(\mathbf{X}) & 0\\ 0 & \epsilon^2\end{pmatrix}$$

and the (two-dimensional!) kde estimator $p^{\epsilon}(x,y)$ for the data $(\mathbf{X},\mathbf{n})$ with product kernel K and the scalling factor $\mathbf{H}=h\cdot\mathrm{cov}(\mathbf{X},\mathbf{n})$ becomes

$$\frac{1}{|\mathbf{X}|}\sum_{x_i,y_i}\int K_{h\sigma^2}(x-x')K_{h\epsilon^2}(y-y')\delta_{x_i,y_i}(x',y')\mathrm{d}(x',y')$$

with $\sigma^2=\mathrm{Var}(\mathbf{X})$ and the summation over $x_i\in\mathbf{X},y_i\in\mathbf{n}$. Note this is supported on $y\in[-\epsilon,\epsilon]$. We now have for every $x\leqslant r$

$$\int_{-\epsilon}^{t-x+\epsilon} p^{\epsilon}(x,y)\mathrm{d}y = \int p^{\epsilon}(x,y)\mathrm{d}y = \frac{1}{|\mathbf{X}|}\sum_{x_i\in\mathbf{X}}\int K_{h\sigma^2}(x-x')\delta_{x_i}(x')\mathrm{d}x'$$

which is a kde approximation for $\mathbf{X}$.

Facilitating Public Access to Legal Information

A Conceptual Model for Developing an Agile Data-driven Decision Support System

Shefali Virkar
Danube University Krems
Krems a. d. Donau, Austria
shefali.virkar@donau-uni.ac.at

Chibuzor Udokwu
Danube University Krems
Krems a. d. Donau, Austria

Anna-Sophie Novak
Danube University Krems
Krems a. d. Donau, Austria

Sofia Tsekeridou
INTRASOFT International S.A.
Peania Attikis, Greece
sofia.tsekeridou@intrasoft-intl.com

***Abstract*— The European legal system is multi-layered and complex, and large quantities of legal documentation have been produced since its inception. This has significant ramifications for European society, whose various constituent actors require regular access to accurate and timely legal information, and often struggle with basic comprehension of legalese. The project focused on within this paper proposes to develop a suite of user-centric services that will ensure the real-time provision and visualisation of legal information to citizens, businesses and administrations based on a platform supported by the proper environment for semantically annotated Big Open Legal Data (BOLD). The objective of this research paper is to critically explore how current user activity interacts with the components of the proposed project platform through the development of a conceptual model. Model Driven Design (MDD) is employed to describe the proposed project architecture, complemented by the use of the Agent Oriented Modelling (AOM) technique based on UML (Unified Modelling Language) user activity diagrams to develop both the proposed platform's user requirements and show the dependencies that exist between the different components that make up the proposed system.**

Keywords—legal information retrieval; data-mining; text-mining; requirements engineering; conceptual-modelling.

I. Introduction

The objective of the project under study [name withheld to preserve anonymity] is to enable access to legal information across the European Union and improve decision-making in legislative procedures through the provision of the appropriate technical environment and the tools for making legal information available to everybody in a customizable, structured and easy-to-handle way. To achieve this, the project seeks to conceptualise and build a suitable environment for semantically annotated Big Open Legal Data (BOLD), one that makes it easily searchable and exploitable with proper visualization techniques.

Accurate, target-orientated, and timely legal information is a strategic input in effective decision-making for a plethora of actors within society. However, the European legal system is multi-layered and complex, and large quantities of legal documentation have been produced, digitized and published online since its inception. This sheer mass of legal information - in the form of legislation, case law, law doctrine, and citizen-generated content - currently remains fragmented across multiple national databases and hidden within inaccessible systems, and is often a challenge for citizens, businesses and legal practitioners alike to access, retrieve and/or comprehend.

To resolve this dilemma, the project proposes the development of the following user-centric suite of services: research through legal corpora, analyzing the alignment of national legislation with EU legislation, comparing national laws which target the same life events, analyzing the references to European legislation by national laws, analyzing related laws within the same Member State, timeline analysis for all legal acts, visualization of the progress and current status of a specific national or European piece of legislation and sentiment analysis towards new legislation.

II. Background and Structure of the Paper

A. Background

The performance and level of acceptance of any information retrieval system that is operated by human actors depend on the correct identification and fulfilment of user requirements, based on the appropriate identification of stakeholders and users. In the Information Systems literature, stakeholders have been broadly defined as all those actors whose actions may be influenced or be impacted directly or indirectly by the use of a given system [1]. Users are one such group of so-called 'baseline stakeholders' – the people, groups, and organisations who directly interact with a system, operate it, and make use of its products and services [2]. Users may be further categorised into different sub-groups and investigated according to either the roles they play within a given social, political or economic context (age, gender, occupation, etc.), or by the nature of their interaction with a given information system (familiarity with IT, patterns of computer use, system expectations, etc.).

Bridging the gap between current usage patterns and optimal usage of a system relies on the mapping out of the available functionalities of the system and their interaction with the operation of the system by users to produce system output. These user actions are defined both by the options conveyed by

The ManyLaws project is co-financed by the Connecting Europe Facility of the European Union, *CEF-TC-2017-3 Public Open Data.*

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_10

the user interface, and by the decisions of the users at the point of input. Correctly identifying users and the roles they play within a given context is hence central to the identification of user requirements based on the construction of accurate user interaction models (UIMs) that map out how all the objects and user actions within a given system interrelate.

B. Structure of the paper

The objective of the authors is to develop a position paper that critically explores how current user activity interacts with the components of the proposed project platform. The main research question for this paper is how to design a legal information system that is capable of delivering an advanced user-generated query service using technologies available in text mining, data mining and big data repositories coupled with open legal data sources. Thus, the first step in the development of a successful conceptual model is user requirements analysis, followed by an investigation of the processes that currently take place and a description of the proposed functionality of the system. The technical core of the paper will be developed in two stages: first, the proposed solution's functional requirements will be mapped to a user activity diagram created from a selected user story. Next, user activity will be mapped onto the proposed static architecture of the platform to produce a critical narrative of what the project can already deliver.

III. Methodology

To develop the conceptual model for the proposed system, a three-fold approach has been adopted. First, we use Model Driven Design (MDD) to describe the system platform. MDD is a system design technique that solves complexity issues in building large-scale systems consisting of several components [3] by capturing all essential system design information in a set of formal or semi-formal models kept consistent by tools [4]. To complement this, the Agent Oriented Modelling (AOM) technique is used to develop the user requirements of the platform. AOM provides a method, centered on the notion of 'agent', to systematically develop the functional requirements of a system [5]. In combining AOM methodology with the final element in our approach, the user story, one of the most popular requirements artefacts in Agile Software Development [6], we aim to construct a behavioural scenario that describes how agents achieve the goals set for the system by performing a sequence of activities. The interaction model of the platform described comprises of the expected user activities on the platform, and this is modelled using UML (Unified Modelling Language) user activity diagrams. The static architecture of the platform comprising of the service layer, component layer and infrastructure layer is modelled using UML component diagram notations. The static architecture, in particular, shows the dependencies that exist between the different components that make up the proposed system.

IV. Conceptual Model Design

In this paper, we present the conceptual model of the proposed legal information retrieval platform. The model represents both the static and interactive architecture of the platform. The interactive components constitute the user goals and user-generated activities on the platform. The static architecture comprises of the service layer, the processing layer, and the infrastructure layer of the platform.

A. Proposed Solution Interactive Architecture: Functional Goals and User Stories

The interactive components of the proposed solution are represented at the upper layer of the conceptual model, and they form the motivational layer and process layer of the platform. First, the motivational layer based on the functional goals of the platform is derived, and followed by the derivation of the process layer based on the user stories on the platform.

Functional Requirements: Identifying the functional requirements of the platform is also important for the development of comprehensive user interaction models. The functional requirement of the platform is represented in the motivational layer in Fig. 1. The functional requirements represented in this model are the expected user goals and objectives that the platform will provide. In this paper, the agent-oriented goal modelling technique is used to represent the user-based functional requirements of the proposed system [5]. As shown in Fig. 1, the top part illustrates the main requirement of the platform while other requirements are represented as sub-goals of the platform.

The main functional goal of the proposed system is to enable end-user access to legal information in an efficient way. This main objective is further broken down into two first level goals. The user may conduct a keyword-based search by setting a textual query or retrieve already-known legal information, presented in an intuitive manner. The first level goals of the platform are used in deriving the service architecture of proposed system. The first level goal Query Keyword Search is linked to the following sub-goals Comparative Analysis and Result Presentation. Under the comparative analysis, the user can perform the following functions: explore the alignment of national laws with EU directives, compare the national laws of different member states, compare related laws within the same member state, or compare references to EU directives within the national laws of member states. Under the Result Presentation, the user can simply view in textual form the retrieved, according to the search query, results. The retrieved results in the Query Keyword Search can also be visualized intuitively. For instance, the user can view Timeline Analysis, Comparative Analysis and Dependency Visualizations of the retrieved results. The other first level functional goal Retrieve Legal Information is also associated with intuitive information visualizations and Comparative analysis sub-goals. The same functions previously stated also apply to these sub-goals.

The proposed solution platform ensures that information is presented in multilingual format, and also provides parallel search capabilities by providing information simultaneously from different national databases. The multilingual capacity of the proposed system is a quality requirement of the platform, and is, therefore, associated to the main functional requirement of the platform which is to provide access to legal information (in multilingual format). The Parallel Search capability of the platform is also defined in this case as a quality requirement of the platform, and is associated with the first level requirement Query Keyword Search.

User Stories: There are many different types of users who are potentially interested in having improved access to legal information. Legal Professionals constitute the most obvious user group, possessing a direct professional requirement for comprehensive, accurate, and timely legal knowledge. An examination of current and ongoing projects in the field of legal information helped project researchers identify three further broad user categories: Citizens, Businesses, and Government. Interviews and focus groups facilitated the construction of 6 initial user stories [7]: Citizen, Lawyer (x2), Business Person, Legal Administrator, Parliamentary Administrator; from which 6 top-level user groups were derived for further study: Citizens, Scientific Community, Legal Professionals, Business Persons, Public Servants, and Third Sector Actors.

User stories are represented in the process layer of the platform. The user activities on the platform are based on the functional requirements of the platform, and are supported by the service layer of the platform. User stories are short, simple descriptions of an information system or product use narrated in natural language from the perspective of an end-user [8]. User stories are considered an integral part of an agile approach to system development that help map usage patterns or strategies involving current resources and tools, and which identify user expectations and functional requirements [9]. User stories, on the one hand, help agile system and product development teams keep track of the needs, expectations, and values of their target user groups, and, on the other hand, help them identify the activities they need to engage in to achieve these goals [10].

The process layer of the conceptual model is derived using the UML activity diagram notations. Within this research paper, a single most indicative user story - Citizen - is represented in the conceptual model. Citizens represent the largest, most diverse user group of information system services, applications, and products. They may be studied in terms of differences in age, gender, education, income, occupation, geographic location, nationality, to name a few – all of which influence their requirement for legal information, their ability to access and understand legalese, their go-to sources of legal information, the search strategies they employ, and their comfort with information technology. They also have the widest need for legal information as they are impacted by different laws at different moments, are entitled to a wide number of legal rights and are bound by numerous legal obligations/duties. Citizens are also highly mobile - as individuals, groups or families they regularly travel across borders for leisure, on work, to seek temporary employment, or to immigrate permanently. As they are not expected to possess a prior specialist knowledge of legal matters, and given the current preference for top-down legislative procedures, citizens are usually only purely consumers of legal information. They are therefore highly dependent on expert legal advice, and generally require assistance with locating and interpreting legal information. An additional consideration is cost, wherein citizens are often not in a position to, or might not be willing to pay, high legal fees.

Citizen User Story: Karolina is a 25-year old Austrian living in the Netherlands. Her user story was first reported in [7] as an initial use case scenario, and excerpts from the interview transcript are reproduced verbatim here:

"As a private citizen living abroad within the European Union, I find that I require legal information both for work and personal use, although I am more concerned about legal information related to personal matters. More particularly, I feel that I need to be familiar with legal information both from my country of residence (the Netherlands) and from my home country (Austria). I recently needed information about the law concerning personal injury claims in the Netherlands. Being Austrian, my car is bought and registered in Austria, and I am concerned that I do not know enough about my legal obligations closer to where I live. I currently look for legal information both online, via Google searches, and offline, using Citizens Advice Bureaus and legal experts…."

"...My first activity would be to use Google for a simple keyword on the topic, and accesses those hits that look like they link to reliable, official sources. I prefer National Open Data portals and government data portals, as they are comprehensive sources of accurate basic legal information. However, I think that these portals lack an easy interpretation of legal language, and they often do not offer pan-European or comparative information. In fact, they do not go much further than the provision of the actual law itself. I do not speak fluent Dutch, so sometimes this is also a problem. If I cannot find all the information I am looking for, or if the information is too complex, I would contact the local citizen advice bureau. If they cannot tell me where to file a complaint, or give me adequate information, I would consider seeking legal advice from a legal expert."

Identified User Needs and Expectations: What stands out from Karolina's user story is the requirement for a comprehensive, authoritative, multilingual online source of legal information. Like many other young, upwardly mobile EU citizens who do not possess specialist legal knowledge, Karolina is concerned that her information comes from authoritative sources. Karolina is also very comfortable with information technology, and tends to use the internet as her primary port of call for information. She hence usually resorts to looking for legal information on official government portals. However, she feels that these portals, while being fairly comprehensive legislative databases, need to offer more in terms of easily comprehensible and comparative pan-European content. Although bilingual, Karolina finds it difficult to contend with European languages that are not her native German or English. Her dependency on external expert interpretation of legal information is, therefore, highlighted.

The following activities outline the expected user behaviour when interacting with or using the platform in search of legal information with regards to making a personal injury claim as described in the user story:

- Activity 1: Access proposed system platform,
- Activity 2: Search for legal information
- Activity 3: View relevant search results in a ranked manner
- Activity 4: Select the most suitable result

- Activity 5: Identify connected laws
- Activity 6: Identify reference to European directives in national laws of member states
- Activity 7: Access visualisations to determine conflicts, conditions, eligibility,
- Activity 8: Compare with relevant law(s) in home country member state,
- Activity 9: Access visualisations to make comparison with home country.

These actions, represented diagrammatically as the process layer in Fig. 1, are described as follows. The user accesses the platform to search for legal information (in this case, to make a personal injury claim). They perform a simple keyword search to retrieve relevant information. The platform offers them a list of relevant hits from across the European Union. The user selects the most relevant result, in this case the relevant national law from the Netherlands. Depending on the expert level of the user, they either are able to interpret the complexity of the law or require further interpretation. In this case, they proceed to the next step by identifying connected laws within the same member state, and also identifying references to European directives within these laws. The user will also be able to view in visual form the conflicts and conditions of eligibility within the given member state. At this point, the user can either decide to exit the platform or get more information concerning the same or similar laws in their home country. The user can then use visualisations to make this comparison. The user can then exit the process.

B. *Static Architecture*

The static architecture provides a high-level description of the components of the proposed system platform, and these are derived using UML modelling notations. At the top of the static architecture is the service layer which represents the user-based services offered by the platform. This is followed by the information processing components. At the bottom of the static architecture component is the infrastructure layer, which consists of the data management layer and the security layer.

Services: The service layer supports the process layer in enabling the users to interact with the platform. The services that make up this layer are directly derived from the first level user requirement of the motivational layer. This comprises of user-generated query service and information retrieval service. The proposed solution services are supported by the Structural Data Component and Visual Analytics Component of the information processing components. The proposed system query search engine will provide to users the capacity to find relevant information efficiently and the results will be presented in a user-friendly manner. The users' searches will be analyzed for optimization purposes. The search component will consider semantic characteristics of the search terms and stored legal resource entity in order to present the most suitable results to the user.

Information processing components: This consists of technologies for processing the data generated from the data management layer. The information processing components support the service layer of the proposed system platform. This comprises the following components; Translation Services, Text Mining, Structured Data and Visual Analytics Components. The structured data component supports the user-generated queries and retrieval service while the visual analytics component supports the intuitive visualization service. The translation services component prepares the data acquired from the big data sources before processing by the text mining tools. The text mining tools use various algorithms supported by a super computing infrastructure and generate intermediate results. The structured data component converts the harvested results into a common structured format such that each element of the data can be effectively processed and analyzed further by the visual analytical component. The visual analytical component converts the structured data into an interactive visual representation.

Infrastructures: The infrastructure layer of the proposed solution conceptual model consists of Data Management Infrastructures and Security Infrastructures. This layer supports the component layer of the system. The data management infrastructures are the data sources and relevant legal artefacts, residing in their original data repositories, that provide inputs for the data processing and mining carried in the proposed system. This is accomplished by interfacing with data access APIs at the source origin, and, after textual mining, extracting the structural and legal-specific semantic metadata stored in the metadata repository of the proposed system. The data sources are mainly national legal databases, and EU legal databases such as EUR-Lex. All these ideally open data sources formulate, in a distributed way, one open law data source, which when accompanied by the universal legal metadata repository described above, formulate the Big Open Legal Database (BOLD). It is noted that the proposed system does not target to aggregate all original legal documents into one single database – it is instead devising a universal legal metadata and index store for all these open data sources, which its upper layer query and search services access to respond to relevant user queries. The BOLD database therefore provides the main input for the translation service component in the component layer of the static architecture. The security infrastructures consists of components and services that are necessary to secure the proposed system platform, The security layer, therefore, supports the other infrastructures, components and services that make up the proposed system. To ensure trust and reliability of data processed in the proposed system platform, the following external infrastructures will be integrated to the platform: Public Key Infrastructure, Digital Rights Management mechanisms, Risk Prediction Algorithms. To ensure that the proposed system platform is protected from unauthorised access and related attacks, the proposed solution implements user management authentication and authorization services.

V. Discussion

Modelling user stories is crucial for agile system development, as it enables system architects to identify and respond quickly to changing user needs [11]. The current user requirements which are represented in the goal model are derived from the proposed list of services which the proposed

solution is set to deliver. Additional study is required to further refine the conceptual model and evaluate end-user behaviour; and, in particular, to determine how different target user categories will interact with the system and use the various functionalities of the proposed solution services.

The conceptual model and functionalities described in the project under consideration, represented schematically in Fig. 1, have a few similarities with a number of existing legal information systems such as OpenLaws [12]. Both projects seek to provide and increase access to legal information. OpenLaws achieves this by building a big open legal database (BOLD) to provide structured legal information to the users [12]. The proposed solution discussed in this paper will also rely on such an approach, accessing the data in BOLD database for big data acquisition. However, the project team seeks to also develop advanced data analytics and text mining solutions, for multilingual resources, that will allow users to pose advanced queries, upon which the system will semantically correlate them with the annotated resources, to allow the user to retrieve relevant and detailed results, visualized in intuitive ways, and thus allowing better understanding of legal information. This is achieved using semantic analysis, comparative and visualized data analytics techniques. Semantic analysis and annotation is achieved using a legal documents ontology mapping the main structural and semantic entities in the form of a graph with relevant relationships. This is further

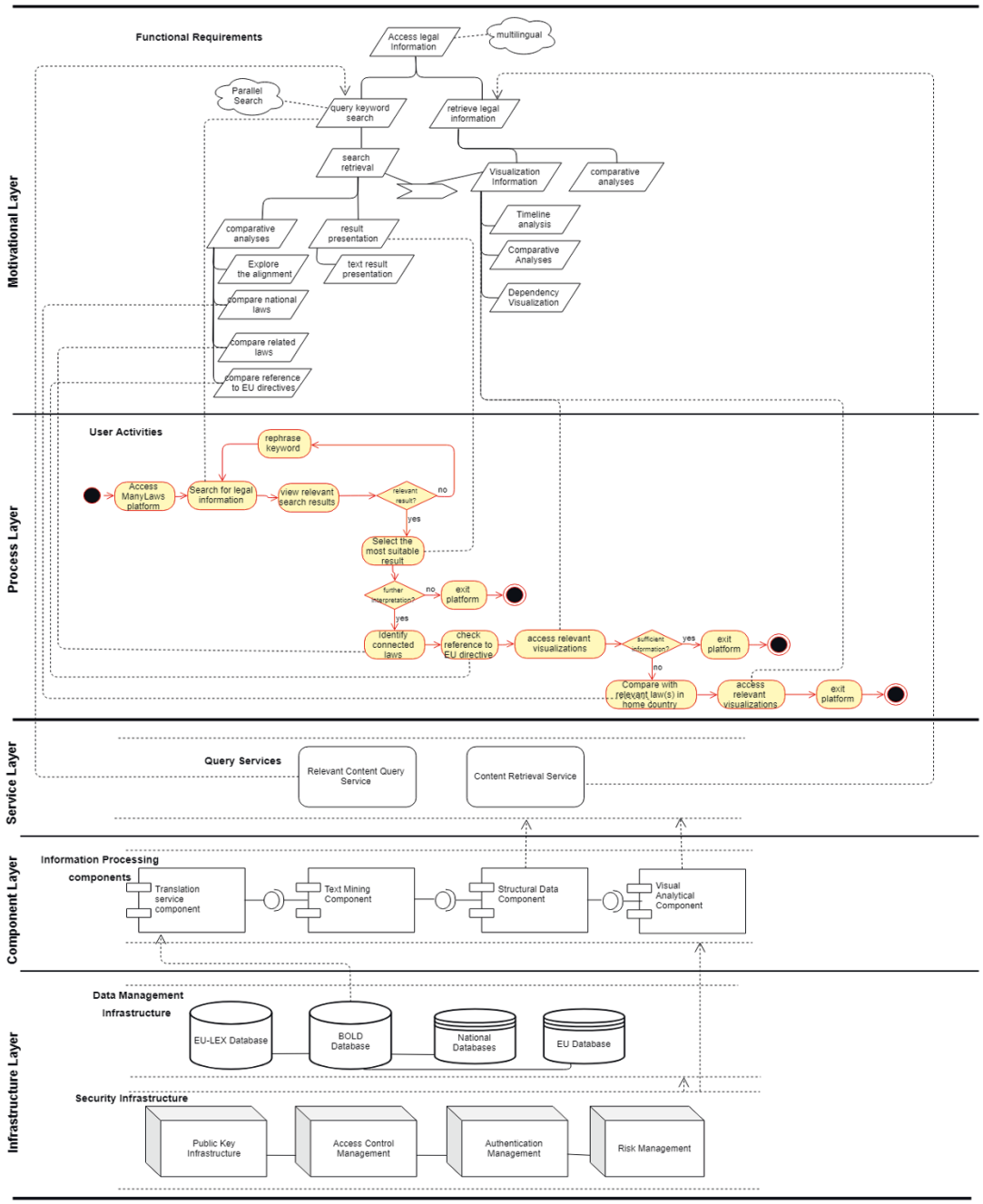


Fig. 1. Proposed System Conceptual Model

complemented by an augmented metadata schema providing the structure and values for the semantic entities, derived and extending existing standard document and legal metadata schemata and ontologies (DCAT, ELI). During the analysis and annotation process, text mining results are reasoned upon the legal ontology – this process may be time-consuming but as it occurs during the processing phase, capitalizing on top of high performance computing infrastructures, it does not pose a constraint.

Agent oriented goal modelling is currently a dominant paradigm in requirements engineering. It is of growing recognition that complex information systems are best described as socio-technical systems which consists of human interactions with automated and semi-automated software components [13]. The focus of goal-oriented requirements engineering is to systematically determine the objectives of the system and also identify the software components and human users that are associated with each objective [5]. In this case, both the software components and human users are referred to as 'agent'. This study only considered system goals which are associated to the users. However, the system goals that are associated with software components are not represented in the conceptual model developed from this study. As a result, using the agent oriented goal modelling technique, we are able to generate the user requirements of the proposed solution platform. The proposed solution, being a complex socio-technical platform consisting of several classifications of human users, multiple components interacting with each other as well an external infrastructure component, can further explore AOM in systematically deriving the functional requirements of these components.

One potential limitation of this research paper could be its focus on a single user group, that of citizens, the decision to model the behaviour and interaction activity of one representative of one target user group. However, as argued previously, opting to model our citizen indicative user story was a strategic research choice. As a category, citizens are not only the most numerous and diverse, they also interact with the law in the broadest possible context, and are the least likely of all identified user groups to possess expert legal knowledge. This implies that they will come to depend the most on the proposed solution platform, use its services more frequently to locate and interpret legal information than other more specialised users, and combine more of the features during each interaction to create value and extract meaning. Anticipating information needs and modelling citizen activity on the proposed solution platform is thus a complex task that needs to be tackled first, not last.

VI. Conclusion

In this study, we developed a conceptual model that describes how to design an information system that is capable of delivering an advanced user-generated query service using text and data mining technologies. The model consists of the static architecture, goal model and user activity. The static architecture contains the external infrastructure, information processing components and services that makes up the platform. The model also shows how the various layers of the model interact and support the other layers.

The proposed system is a user-centric platform, therefore it is necessary to develop the functional requirements of the system by first focusing on user requirements. The agent oriented goal modelling approach provides a method for systematically deriving the user requirements based on the main objectives of the platform. The logical next step in this process is, therefore, to model all initial user stories based on available interview transcripts. These models can be used to identify user needs, priorities and information-seeking strategies which, in turn, may be applied to the development of user requirements elicitation questionnaires and interviews. The data obtained from these research tools can be utilised to improve existing project offerings and to develop additional services through the definition and refinement of functional and non-functional requirements and subsequently the detailed system architecture and design specification.

References

[1] A. Pouloudi, and E.A. Whitley, "Stakeholder identification in inter-organisational systems: gaining insights for drug use management systems," European J. of Inf. Sys., vol. 6, iss.1, pp. 1-14, 1997.

[2] H. Sharp, A. Finkelstein, and G. Galal, "Stakeholder identification in the requirements engineering process," in Proceedings - Tenth International Workshop on Database and Expert Systems Applications. DEXA 99, IEEE Press, 1999, pp. 387-391.

[3] K. Balasubramanian, A. Gokhale, G. Karsai, J. Sztipanovits, and S. Neema, "Developing applications using model-driven design environments," Computer, vol. 39, iss. 2, pp.33-40, February 2006.

[4] A. Mattsson, B. Lundell, B. Lings, and B. Fitzgerald, "Linking model-driven development and software architecture: A case study," IEEE Tran. on Softw. Eng., vol. 35, iss.1, pp. 83-93, 2009.

[5] L. Sterling and K. Taveter. The Art of Agent-Oriented Modeling. Cambridge, M.A.: MIT Press, 2009.

[6] B. Ramesh, L. Cao, and R. Baskerville, "Agile requirements engineering practices and challenges: an empirical study," Inf. Sys. Journal, vol. 20, iss. 5, pp. 449-480, 2010.

[7] Y. Charalabidis, M.A. Loutsaris, S. Virkar, C. Alexopoulos, A.-S. Novak, and Z. Lachana, "Use case scenarios on legal machine learning," in Proceedings - ICEGOV 2019, in press.

[8] M. Cohn, User Stories Applied: For Agile Software Development. Boston, M.A.: Addison-Wesley Professional, 2004.

[9] M. Düchting, D. Zimmermann, and K. Nebe, "Incorporating user centered requirement engineering into agile software development," in Proceedings - International Conference on Human-Computer Interaction, Berlin, Heidelberg: Springer, 2007, pp. 58-67.

[10] T.S. Da Silva, A. Martin, F. Maurer, and M. Silveira, "User-centered design and agile methods: a systematic review," in Proceedings - Agile Conference (AGILE) 2011, IEEE Press, 2011, pp. 77-86.

[11] F. Paetsch, A. Eberlein, and F. Maurer, "Requirements engineering and agile software development," in Proceedings - Twelfth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, 2003. WET ICE 2003. IEEE Press, 2003, pp. 308-313.

[12] T. J. Lampoltshammer, C. Sageder, and T. Heistracher, "The openlaws platform—An open architecture for big open legal data," in Proceedings of the 18th International Legal Informatics Symposium IRIS, vol. 309, 2015, pp. 173-179.

[13] A. Van Lamsweerde, "Goal-oriented requirements engineering: A guided tour," in Proceedings - Fifth IEEE International Symposium on Requirements Engineering, 2001. IEEE Press, 2001, pp. 249-262.

Do we have a Data Culture?

Wolfgang Kremser
Salzburg Research Forschungsgesellschaft m.b.H.
Salzburg, Austria
wolfgang.kremser@salzburgresearch.at

Richard Brunauer
Salzburg Research Forschungsgesellschaft m.b.H.
Salzburg, Austria

***Abstract*—Nowadays, adopting a "data culture" or operating "data-driven" are desired goals for a number of managers. However, what does it mean when an organization claims to have data culture? A clear definition is not available. This paper aims to sharpen the understanding of data culture in organizations by discussing recent usages of the term. It shows that data culture is a kind of organizational culture. A special form of data culture is a data-driven culture. We conclude that a data-driven culture is defined by following a specific set of values, behaviors and norms that enable effective data analytics. Besides these values, behaviors and norms, this paper presents the job roles necessary for a data-driven culture. We include the crucial role of the data steward that elevates a data culture to a data-driven culture by administering data governance. Finally, we propose a definition of data-driven culture that focuses on the commitment to data-based decision making and an ever-improving data analytics process. This paper helps teams and organizations of any size that strive towards advancing their – not necessarily big – data analytics capabilities by drawing their attention to the often neglected, non-technical requirements: data governance and a suitable organizational culture.**

Keywords—data culture; big data; data-driven culture; data governance; big data analytics

I. Introduction

The term "data culture" repeatedly appears in (online) publications that are thematically close to business analytics and big data. While scholars repeatedly emphasize its significance for a successful, long-term data-centric venture [1][2], they do not provide a clear definition. We believe data culture to be insufficiently defined relative to its already recognized importance. What distinguishes organizations that follow a data culture is vague and "there are very few attempts made to really define what data culture means" [3, para. 2].

This paper aims to provide a definition of data culture by identifying its set of shared values as well as the corresponding job roles and their responsibilities towards a healthy data culture. It concludes by offering a definition centered on data quality and the associated job role of the data steward. This definition makes it clearer in which ways a healthy data culture benefits data analytics processes. Managers will understand the commitment and talent required to establish a fruitful data culture. The job role descriptions will help data analytics professionals to focus their efforts and communicate their needs.

A. How "Data Culture" is used

We start our exploration of the meaning of "data culture" by looking at examples of how it is used in web and literature. Three examples have been selected because they emphasize different aspects and will serve as a guide towards a comprehensive definition:

For Microsoft, a good data culture is one in which "every team and every individual is empowered to do great things because of the data at their fingertips" [4, para. 6]. This description puts strong emphasis on the democratization of data. Empowering every individual within an organization to extract value from data requires a robust set of tools and methods.

A whitepaper by the IT service provider Cognizant describes data culture as a "workplace environment that employs a consistent, repeatable approach to tactical and strategic decision-making through emphatic and empirical data proof" [5, p. 2]. Again we notice an emphasis on robust tools and methods to make data analytics consistent and repeatable. Furthermore, it links data culture to a form of decision making based on data.

Marr says that "in a data culture, data is recognized as a key business asset, and it is used, wherever possible, at every level of the business to make improvements" [6, p. 172]. This description corroborates the democratization and decision making aspects of data culture. It also introduces the aspect of constant improvement associated with data culture.

All three attempts to characterize data culture have in common that data is considered as an asset of an organization and that organizations either "have" or "do not have" a data culture. Nevertheless, they do not outline what a "culture" exactly is: Is a (data) culture a set of common values, a common agreement, a set of recommendations, a way of structuring an organization or a set of business processes?

In their recent systematic literature review about big data capabilities, Mikalef et al. [7] used the term "data-driven culture" instead of "data culture". They classified data-driven culture as an intangible resource within the managerial framework of resource-based theory. A resource is defined as "stocks of available factors that the organization owns or controls" [8, p. 35]. Other examples for intangible resources are organizational culture and organizational learning [9]. Mikalef et al. [7] extracted three aspects that contribute towards data-driven culture: (1) prioritization of business analytics investments, (2) support from top management, and a (3) fact-based and learning culture.

B. Data Culture vs. Data-Driven Culture

Before defining data culture, we first want to place the term ontologically. Starting off the considerations of Mikalef et al., data culture is an intangible resource of organizations [7]. This

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_11

puts it on the same level as organizational culture [9]. Indeed, scholars treat data culture as a special form of organizational culture [10]. While there exists no generally accepted definition of organizational culture [11], we use the following working definition: An organizational culture is the collection of common values, behaviors and norms which are postulated by top management and are followed by most individuals within an organization. Data culture therefore refers to the values, behaviors and norms shared by most individuals within an organization with regards to data-related issues.

Under this lens, data culture becomes a very broad term. All modern organizations collect some form of data (e.g. billing addresses, transaction data, supplier data, etc.). Hence all modern organizations have a data culture. The term is too general when trying to talk about specific data-related capabilities of an organization. What is more often meant when talking about data culture is data culture of a certain quality. We call a data culture of this quality data-driven. This differentiation fits the usage of the term "data-driven culture" in [8]. The following sections will isolate the distinct values, behaviors and norms that make a data culture a data-driven culture.

II. Towards Defining Data-driven Culture

As established in the last section, there are certain qualities that distinguish a data-driven organizational culture. To understand these qualities we examine data-driven culture's relation to other organizational resources.

A. Organizational Values

As already stated, a data-driven culture is a form of organizational culture. As such it represents in our understanding a set of values etc. We stipulate three values which have to be shared by most individuals following a data-driven culture. These values can be summarized by an organization-wide preference for data-based decisions:

Value 1 – data-driven decision making is valuable: Decision making on the basis of insights extracted from data is actively encouraged throughout the whole organization. This behavior is favored compared to the "classic" decision-making process in which the hierarchically highest person decides alone and based on experience, costs and intuition [10][12].

Value 2 – seamless data access and data provisioning between business units is valuable: The previous value goes hand in hand with the democratization of data throughout the whole organizational structure. Sharing data between business units enhances isolated data and provides decision-makers on all levels with a basis for their decisions [13].

Value *3 – discussing and distributing knowledge about data processing is valuable*: As a consequence to the distribution of decision-power, a data-driven culture embraces organizational learning which is the process of creating, retaining, and transferring knowledge within an organization [14]. The tools available for manipulating data are numerous, with new solutions for sub-processes being released constantly [15]. This requires professionals to regularly update their knowledge and methods. They consistently share their skills and practices with others to improve data-related processes.

B. The Relation to (Big) Data

One resource related to data-driven culture (DDC) is data itself. As a tangible resource, data cannot create any business value on their own. They lie dormant until they are actively leveraged [9]. Barring directly selling data, the method to generate business value, i.e. insight, from them is called *data analytics* [16].

Data culture has emerged as a topic of discussion in the context of big data and big data analytics. At least since the early 2010's, the term big data has found its way into literature. It is often characterized by the three V's – volume, variety and velocity [17]. De Mauro et al. [18] include big data's requirements on methodology: "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value" [18, p. 131]. These definitions all include the size of the datasets that are processed.

Others define big data without referencing the size of the datasets. The MIKE 2.0 project characterizes big data by the degree of complexity that datasets possess and the amount of value that is gained by applying innovative techniques in analyzing them [19]. Gantz and Reinsel [20] state that big data focuses on three main characteristics: the data itself, the analytics of the data, and presentation of the analytics' results that allow the creation of business value in terms of new products or services.

Whichever definition one chooses, there is no reason why the values listed earlier could only be present in organizations that process big data (BD). Data-driven decisions can be made on the basis of simple and small, yet representative datasets. These datasets can just as well be made accessible, and organizational learning is not bound to any specific kind of dataset. This suggests that a data-driven culture can be established regardless of the size or complexity of the datasets involved. Hence, processing big data cannot be a necessary requirement for data-drive culture (formally: DDC $\not\Rightarrow$ BD).

There is also no reason to believe that all organizations that perform (big) data analytics follow a data-driven culture. It is possible to generate insights from data without ever using these insights as a decision basis, or while keeping data isolated within business units. Simply having data analytics capabilities (DAC) is not enough to claim data-driven culture (DAC $\not\Rightarrow$ DDC).

Instead, it seems that data-driven culture is more closely related to *data quality* (DQ) because "the most significant benefit of quality data is the willingness of managers to trust the data and to act on the information" [21, p. 5]. Therefore we include data quality as a necessary (and reasonable) prerequisite for data-driven culture (DDC $\Rightarrow$ DQ). Data is of high quality if they meet the expectations and needs of the job role that is working with them [22]. The Strong-Wang framework [23] describes 15 dimensions of data quality which later have been summarized into six core dimensions [22]: completeness, uniqueness, timeliness, validity, accuracy, and consistency. Furthermore, in an ongoing effort to standardize data quality, the ISO 8000 [21] defines its own five key characteristics of quality data. One of them, *provenance*, is knowing the source of the data to establish trust.

C. Roles in a Data-driven Organization

For the sake of establishing data quality, and consequently data-driven culture, it is important to know the different job roles involved, together with their perspectives, interests and expectations on data.

First, building and maintaining a data-driven culture needs a strong, organization-wide commitment that starts at the uppermost hierarchical positions (cf. [7]). The *enabler role* of the top-managerial staff is to embrace and propagate the values of a data-driven culture. In order to be able to make better decisions, they have to sufficiently invest in building, maintaining and improving all aspects of data processing (cf. [7]). Without this commitment, organizations will see less results from their data analytics [10], and the daily work between different business units will be cumbersome for data analytics professionals.

Second, the *analytics roles*: To generate insights, data needs to be actively leveraged using data analytics. Facilitating data analytics requires definition of assignment of specific job roles. De Mauro et al. [24] conducted a systematic screening of job offerings in the field of big data analytics in the United States. While they write exclusively about big data analytics, we adopt their job role descriptions for a general data analytics process. In less complex settings, multiple job roles might be assigned to a single person. However, this does not change the overall process. The authors suggest a natural clustering of data-related skillsets into two large groups, *technology-enabler professionals* and *business-impacting professionals*.

Technology-enabler professionals might be considered the backend of (big) data analytics. They are developers and engineers tasked with providing the necessary toolset for data analytics. They are further distinguished into *data engineers* who are responsible for building and maintaining the technology infrastructure (storage, access and processing of data), and *data developers* who build data-reliant applications on top of this infrastructure, e.g. dashboards. [24]

The second group, Business-Impacting professionals, perform the data analysis to create organizational value. They as well are split into two specialist groups: the *data scientists* who are extracting information from the data they are provided, and *business analysts* responsible for project management and translating information into actual business policy. [24]

The roles within the data analytics team have a strong dependency on each other: Data developers depend on the data engineers' infrastructure for their applications which in turn are used by the data scientists to explore the data and derive knowledge. The data scientists communicate their findings to the business analysts who act upon the new found knowledge.

However, nobody in this set of roles directly works towards the values of a data-driven culture. Who coordinates the seamless provisioning of data? Who manages data quality issues to ensure others are willing to rely on data to make their decisions? Who monitors the data-related processes to improve upon them? Who has a holistic and high-level overview?

Therefore it is necessary to add a fifth, essential role to data-driven culture: the *data steward* [25]. While the analytics roles work towards data analytics capabilities, and the enabler role provides managerial direction and support, the data steward has "accountability and responsibility for data and processes that ensure effective control and use of data assets" [22, p. 75]. They have a strong relationship with both the enabler role and the analytics roles due to them being involved in the entirety of the data analytics process.

One major task is ensuring that the other roles are effectively communicating their needs. Since data quality depends significantly on how well the data fit another job role's expectations, it is crucial for the data steward to learn them. Without being asked, data analytics professionals might work around their low-quality data without articulating what is bothering them. [22]

Other tasks include [22]: (1) creating and managing metadata, (2) documenting rules and standards around data, and (3) executing operational data governance activities. The next section explains why data governance in particular is a necessary condition for a data-driven culture, and therefore justifies the inclusion of the data steward as an essential job role.

D. Structures in a Data-Driven Organization

To accommodate a data-driven culture, an organization needs a horizontal structure across job roles and business units. This structure allows the enabler role to promote their values and supports the analytics roles in their daily work.

It seems that data governance provides exactly this horizontal structure. Data governance is classified as an intangible resource [7] and it is recognized to be an important factor for successful data analytics [1][26]. Seiner [25, p. 2] defined data governance as "the formal execution and enforcement of authority over the management of data and data-related assets". With this understanding, data governance is the missing structural asset that supports all key players in an organization with a data-driven culture.

More concretely, data governance is involved in all processes, policies, organizational structures and developments that concern data. It ensures availability, usability, integrity, consistency, auditability, and security of data. In other words, the whole data analytics process, from raw data to extracted insight, is happening in a controlled environment created by data governance. As a consequence, data quality becomes transparent and efficiency becomes measurable. Furthermore, compliance with ethical and legal regulations is checked by data governance. [22]

Without data governance, organizations face a number of risks. Resources like time and money can be wasted in inefficient data analytics without anyone noticing. Data analytics professionals might become frustrated by wrongly assigning them data governance tasks which should go to the data steward instead. The lack of data governance also means that the legitimacy of using data is questionable at best. This can cause severe public relation and legal problems. [22]

And lastly, directly related to data-driven culture, data quality becomes opaque. Decision makers hesitate to base their decisions on data they do not know the quality of, so they might fall back on using intuition, experience and cost. It is clear now

that data governance must be part of a data-driven culture in a significant manner.

E. A Stipulative Definition of Data-Driven Culture

Based on the discussion in the previous sections, we propose the following informal definition for data-driven culture:

A *data-driven culture* is a specific form of organizational culture in which (1) decisions on all organizational levels are made preferably on the basis of insights extracted from well-governed data; (2) all members encourage organization-wide, seamless access and provision of well-governed data, and (3) they take part in continuous distribution of knowledge and skills about data processing.

This definition stipulates the three values as values of the organizational culture together with well governed data. Thus, well-governed data (WGD) are a necessary condition for a data-driven culture (DDC ⇒ WGD). In other words: To follow these values and having well-governed data means following a data-driven culture.

Our definition allows the inclusion of additional values as long as they are in accordance with the ones that are already proposed. Note that this definition does not bind data-driven culture to the size or complexity of the organization's datasets. Instead, with the inclusion of data governance, it requires data quality to be transparent to the decision makers.

F. The Necessary Enabler for Data-Driven Culture

The definition above contains a final important aspect – the limitation to "well-governed" data. Data governance has to reach a certain level of maturity before it can support the values of a data-driven culture. Seiner [25] defines five levels of data governance maturity:

1. Initial Level: Processes are undocumented and subject to ad-hoc changes demanded by users or events.
2. Repeatable Level: Best practices are established which make results more consistent. However, process discipline is not yet rigorous and there is no data governance institution within the organization.
3. Defined Level: Defined and documented standard processes exist which are regularly audited and improved upon. These processes are used to establish consistency of performance.
4. Managed Level: Process metrics are used by management to control and improve data-analytics capabilities. Processes can be used across projects with minimal need for adaption. End-user staff is able to use tools to learn what data exists within the organization and how it is structured.
5. Optimizing Level: Process performance is continually improved through incremental and innovative changes in technology.

At level 4, well-established processes around data exist that ensure the transparency of data-quality. Tools and applications for data analytics are robust; their functionality and use are documented and actively taught across business units. It is at this level that all staff members, including those not central to data analytics, can potentially access all data available within the organization.

Therefore we regard level 4 to be the lowest level to have well-governed data. At this stage, according to Seiner [25], the data stewards are involved in some capacity in all development efforts. He goes on to say that only five to ten percent of organizations have reached this level, which would imply that it is a rather strict barrier to entry into data-driven culture. This assessment is from 2015 and newer numbers are not available to the best of our knowledge. It would be interesting to survey the current state of organizations' data governance maturity and whether their awareness of data quality issues has increased in recent years.

III. Conclusion

We have investigated the term "data culture" and explored how it is conceived in literature. While its meaning is very broad, it has been used synonymously with data-driven culture. We clearly distinguished data-driven culture as a specific form of data culture and organizational culture. By identifying the values, behaviors and roles of a data-driven culture we were able to state that (1) data-driven culture is not only adoptable in big data analytics, (2) big data analytics capabilities do not imply data-driven culture, and (3) data governance is an often overlooked, non-technical, structural requirement for data-driven culture. Finally, we have proposed a stipulative definition of data-driven culture that includes the values which were identified during the investigation. At its core are data-based decision making and mature data governance.

Acknowledgements

This work was partly funded by the Austrian Federal Ministry for Transport, Innovation and Technology, the Austrian Federal Ministry for Digital and Economic Affairs, and the federal state of Salzburg.

References

[1] G. Cao and Y. Duan, "A Path Model linking Business Analytics, data-driven Culture, and competitive Advantage," ECIS 2014 Proc., Jun. 2014.

[2] S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz, "Big data, analytics and the path from insights to value," MIT sloan Manag. Rev., vol. 52, no. 2, p. 21, 2011.

[3] "Data Culture. What is it and how do we enhance it? – Bloor Research." [Online]. Available: https://www.bloorresearch.com/2018/03/data-culture-enhance/. [Accessed: 20-Jan-2019].

[4] "A data culture for everyone - The Official Microsoft Blog," 2014. [Online]. Available: https://blogs.microsoft.com/blog/2014/04/15/a-data-culture-for-everyone/. [Accessed: 20-Jan-2019].

[5] P. Ramaswamy, "How to Create a Data Culture," Cognizant 20-20 insights, 2015. [Online]. Available: https://www.cognizant.com/InsightsWhitepapers/how-to-create-a-data-culture-codex1408.pdf.

[6] B. Marr, Data Strategy: How to Profit from a World of Big Data, Analytics and the Internet of Things. Kogan Page, 2017.

[7] P. Mikalef, I. O. Pappas, J. Krogstie, and M. Giannakos, "Big data analytics capabilities: a systematic literature review and research agenda," Inf. Syst. E-bus. Manag., vol. 16, no. 3, pp. 547–578, 2018.

[8] P. J. H. Schoemaker and R. Amit, “Strategic Assets and Organizational Rent,” Strateg. Manag. J., vol. 14, no. 1, pp. 33–46, 1993.

[9] R. M. Grant, “The Resource-Based Theory of Competitive Advantage: Implications for Strategy Formulation,” Calif. Manage. Rev., vol. 33, no. 3, pp. 114–135, Apr. 1991.

[10] M. Gupta and J. F. George, “Toward the development of a big data analytics capability,” Inf. Manag., vol. 53, no. 8, pp. 1049–1064, 2016.

[11] J. B. Barney, “Organizational Culture: Can It Be a Source of Sustained Competitive Advantage?,” Acad. Manag. Rev., vol. 11, no. 3, pp. 656–665, 1986.

[12] J. W. Ross, C. M. Beath, and A. Quaadgras, “You may not need big data after all,” Harv. Bus. Rev., vol. 91, no. 12, p. 90--+, 2013.

[13] P. Mikalef, V. Augustin Framnes, F. Danielsen, J. Krogstie, and D. Håkon Olsen, “Big Data Analytics Capability: Antecedents and Business Value,” Twenty First Pacific Asia Conf. Inf. Syst., p. 13, 2017.

[14] L. Argote and E. Miron-Spektor, “Organizational Learning: From Experience to Knowledge,” Organ. Sci., vol. 22, no. 5, pp. 1123–1137, Oct. 2011.

[15] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” J. King Saud Univ. - Comput. Inf. Sci., vol. 30, no. 4, pp. 431–448, 2018.

[16] A. McAfee and E. Brynjolfsson, “Big Data: The Management Revolution,” Harv. Bus. Rev., vol. 90, p. 60–66,68,128, 2012.

[17] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, Apr. 2015.

[18] A. De Mauro, M. Greco, and M. Grimaldi, “A formal definition of Big Data based on its essential features,” Libr. Rev., vol. 65, no. 3, pp. 122–135, 2016.

[19] “Big Data Definition - MIKE2.0, the open source methodology for Information Development.” [Online]. Available: http://mike2.openmethodology.org/wiki/Big_Data_Definition. [Accessed: 18-Jan-2019].

[20] J. Gantz and D. Reinsel, “The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east,” IDC iView IDC Anal Futur., pp. 1–16, 2012.

[21] P. R. Benson, “ISO 8000 Quality Data Principles,” ECCMA, 2019. [Online]. Available: https://eccma.org/private/download_library.php?mm_id=22&out_req=SVNPIDgwMDAgUXVhbGl0eSBEYXRhIFByaW5jaXBsZXM=.

[22] DAMA International, DAMA-DMBOK, 2nd ed. Technics Publications, 2017.

[23] R. Y. Wang and D. M. Strong, “Beyond Accuracy: What Data Quality Means to Data Consumers,” J. Manag. Inf. Syst., vol. 12, no. 4, pp. 5–33, 1996.

[24] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, “Human resources for Big Data professions: A systematic classification of job roles and required skill sets,” Inf. Process. Manag., vol. 54, no. 5, pp. 807–817, Sep. 2018.

[25] R. S. Seiner, Non-invasive data governance: the path of least resistance and greatest success. Technics Publications, 2014.

[26] A. B. Posavec and S. Krajnovic, “Challenges in adopting big data strategies and plans in organizations,” 2016 39th Int. Conv. Inf. Commun. Technol. Electron. Microelectron. MIPRO 2016 - Proc., pp. 1229–1234, 2016.

Non reviewed short Papers

Neural Machine Translation from Natural Language into SQL with state-of-the-art Deep Learning methods

Dejan Radovanovic
Salzburg University of Applied Sciences
Puch bei Hallein, Austria
dejan.radovanovic@fh-salzburg.ac.at

***Abstract*—Reading text, identifying key ideas, summarizing, making connections and other tasks that require comprehension and context are easy tasks for humans but training a computer to perform these tasks is a challenge. Recent advances in deep learning make it possible to interpret text effectively and achieve high performance results across natural language tasks. Interacting with relational databases trough natural language enables users of any background to query and analyze a huge amount of data in a user-friendly way. This paper summaries major challenges and different approaches in the context of Natural Language Interfaces to Databases (NLIDB). A state-of-the-art language translation model developed by Google named Transformer is used to translate natural language queries into structured queries to simplify the interaction between users and relational database systems.**

***Index Terms*—Natural Language Interface to Databases, Distributed Word Representations, Neural Machine Translation, Transformer**

I. Introduction

With the growth of mobile applications and connected systems, the amount of data produced daily is increasing exponentially. Most of this data is stored in relational databases and extracting useful information from those databases requires specific knowledge and familiarity with query languages like SQL (Structured Query Language). This is the reason why natural language interfaces have received attention as helpful tools, which simplify the interaction between users and computers. Part of the research in the field of NLP revolves around the interfacing of databases. The Natural Language Interface to Databases (NLIDB) allows users to retrieve useful information from any database by removing the need to know the specifics of database query languages. Constructing an interface like this requires understanding the context of natural language sentences as well as converting these sentences into meaningful executable queries, be it logical forms or SQLs.

The main difficulties in designing this kind of natural language interface derive from linguistic issues such as language ambiguities and the challenge to develop a general-purpose solution that is portable to different databases. Researchers use different approaches to deal with these problems. This led to the use of semantic interpretation and syntactic parsing to understand users' questions as shown in [1] and [2].

The success of deep learning networks now enables a variety of new possibilities for natural language interfaces to databases. Training supervised neural models from query-answer pairs to retrieve possible answers, is one approach proposed by [3]. A more advanced application is the direct translation of natural language queries to SQL language with machine translation models [4]. In Section II we discuss the idea behind distributed word representations and their importance in machine translation models, which are commonly based on sequential neural networks. The Transformer, a state-of-the-art neural machine translation model that was first proposed in [5], is then discussed in Section III.

II. Distributed Word Representation (Word Embedding)

Recognizing similarities between words is an easy task for the human mind but training a computer to perform this task is a challenge that cannot be overcome by traditional machine learning methods. Experts are crafting features by hand to calculate the representation of words which are then fed into simple neural networks. Developing a multilingual system requires manually crafted features designed by experts fluent in all supported languages. The scalability, maintainability and portability to new languages of systems like these are severely lacking. Deep learning techniques on the other hand offer a way for a network to learn representations on its own. Instead of relying on expert knowledge, this concept uses automatically generated task-independent features (word embeddings) from large amounts of plain text. Word embeddings, a form of distributed word representations, map the index of a word in a dictionary to a feature vector in a high-dimensional space. Every single dimension of this feature space varies in complexity and describes multiple concepts [6]. Using these features as an input for a deep neural network now offers the scalability, maintainability and portability to new languages that hand-crafted features could not provide. Another advantage is the possibility to pre-train these embeddings for large text collections shown by industry standards like word2vec [6]. A novel way to process word embeddings is the aforementioned Transformer, which is discussed in detail in the next section.

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_12

III. Transformer for Neural Machine Translation

The Transformer was introduced by Google in 2017 in the paper ”Attention is all you need” [5]. The transformer is a state-of-the-art Neural Machine Translation (NMT) model that uses self-attention to improve the performance of certain translation tasks. Bevor analyzing the Transformer lets look at the overall challenges in neural machine translation tasks. NMT is an end-to-end learning approach for automated translations that maps single words (sequences) from one language into sequences of another one. Performing such a mapping is done with sequence-to-sequence methods that use encoder-decoder models. The encoder takes the input sequence and maps it into a higher dimensional space. This vector is fed into the decoder which turns the intermediate higher dimensional representation into an output sequence. Bevor the Transformer was introduced a commonly used approach in sequence-to-sequences models was to use recurrent networks. Recurrent networks like RNNs (Recurrent neural network), LSTMs (Long short-term memory) and GRUs (Gated recurrent unit) are based on a recurrent construction and handle word-by-word sequentially. The challenge with this recurrent encoder-decoder architecture is that especially long sentences with long-range dependencies get difficult to learn and make it more difficult to take advantage of modern fast computing devices such as TPUs and GPUs [5].

In contrast, the Transformer use the entire sentence and selectively extract the information it needs during the decoding process. It applies a self-attention mechanism which directly models relationships between all words in a sentence, regardless the respective position of the word. Lets consider the following example ”I arrived at the bank after crossing the river”. The Transformer is able to analyze the context of these sentence and determine that the word ”bank” refers to the shore of a river and not a financial institution. Computing the next representation for a given a word - e.g. ”bank” the Transformer compare it to every other word in the sentence and the result of these comparisons is saved in an attention score. These attention scores determine how much each of the other words should contribute to the next representation of ”bank”. In our example the word ”river” could receive a high attention score when computing a new representation for ”bank”. All attention scores are then use as weights for a weighted average representation of all words in the sentence. This average representation is fedded into a fully-connected network to generate a new representation for ”bank”, reflecting that the sentence is talking about a river bank. Figure 1 visualizes the attention as lines connecting the position being updated (left) with the position being attended to (right) for this example. All the different colors identify the corresponding attention heads and the line thickness reflects the attention score. For the translation of the example the tensor2tensor [7], an extension for TensorFlow with a pretrained model (WMT 2014 English-to-German) and the corresponding pretrained weights were used. The experiments in [5] and [7] show that

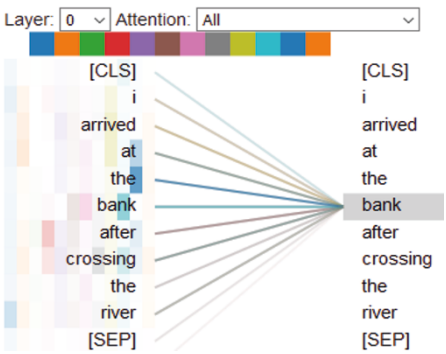


Fig. 1. Visualized attention with tensor2tensor for the example ”I arrived at the bank after crossing the river”.

the Transformer is a high performant sequence-to-sequence model that uses self-attention to increase the accuracy while being more parallelizable and requiring significantly less time to train.

IV. Outlook

Using the Transformer to translate input sequences from English to German was an important task to understand the usage of this network and to deal with overall challenges in neural machine translation. In the next step the Transformer model is going to be trained with an annotated dataset consisting of natural language query and corresponding SQL snippet. The selected data for this step is described in [8]. Another challenge that must be considered for later stages is how database architectures and especially relations between different tables can be used for precise translations.

References

[1] A.-M. Popescu, A. Armanasu, O. Etzioni, D. Ko, and A. Yates, “Modern natural language interfaces to databases: Composing statistical parsing with semantic tractability,” 01 2004.

[2] F. Li and H. V. Jagadish, “Constructing an interactive natural language interface for relational databases,” *Proceedings of the VLDB Endowment*, vol. 8, pp. 73–84, 09 2014.

[3] P. Yin, Z. Lu, H. Li, and B. Kao, “Neural enquirer: Learning to query tables,” *CoRR*, vol. abs/1512.00965, 2015. [Online]. Available: http://arxiv.org/abs/1512.00965

[4] S. Iyer, I. Konstas, A. Cheung, J. Krishnamurthy, and L. Zettlemoyer, “Learning a neural semantic parser from user feedback,” *CoRR*, vol. abs/1704.08760, 2017. [Online]. Available: http://arxiv.org/abs/1704.08760

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” *CoRR*, vol. abs/1706.03762, 2017, arXiv:1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762

[6] R. Al-Rfou, B. Perozzi, and S. Skiena, “Polyglot: Distributed word representations for multilingual NLP,” *CoRR*, vol. abs/1307.1662, 2013. [Online]. Available: http://arxiv.org/abs/1307.1662

[7] A. Vaswani, S. Bengio, E. Brevdo, F. Chollet, A. N. Gomez, S. Gouws, L. Jones, L. Kaiser, N. Kalchbrenner, N. Parmar, R. Sepassi, N. Shazeer, and J. Uszkoreit, “Tensor2tensor for Neural Machine Translation,” *CoRR*, vol. abs/1803.07416, 2018. [Online]. Available: http://arxiv.org/abs/1803.07416

[8] F. Brad, R. Iacob, I. Hosu, and T. Rebedea, “Dataset for a neural natural language interface for databases (NNLIDB),” *CoRR*, vol. abs/1707.03172, 2017. [Online]. Available: http://arxiv.org/abs/1707.03172

Smart recommendation system to simplify projecting for an HMI/SCADA platform

Sebastian Malin
FH Vorarlberg
Dornbirn, Austria
sebastian.malin@fhv.at

Kathrin Plankensteiner
FH Vorarlberg
Dornbirn, Austria

Robert Merz
FH Vorarlberg
Dornbirn, Austria

Reinhard Mayr
COPA-DATA GmbH
Salzburg, Austria

Sebastian Schöndorfer
COPA-DATA GmbH
Salzburg, Austria

Mike Thomas
COPA-DATA GmbH
Salzburg, Austria

***Abstract*—Modelling and connecting machines and hardware devices of manufacturing plants in HMI/SCADA software platforms is considered time-consuming and requires expertise. A smart recommendation system could help to support and simplify the tasks of the projecting process. In this paper, supervised learning methods are proposed to address this problem. Data characteristics, modelling challenges, and two potential modelling approaches, one-hot encoding and probabilistic topic modelling, are discussed.**

The methodology for solving this problem is still in progress. First results are expected by the date of the conference[1].

I. Introduction

In modern industry, HMI/SCADA software platforms are state-of-the-art for computer-aided monitoring and controlling of automated manufacturing processes. Representing a vast number of variables, linked to sensors and actuators from a variety of different machines in a uniform data model is an important feature of these systems. The current practice to manually enter the variables, each consisting of metadata like a unique name, data type, data source and address information is considered time-consuming and expensive. Large automotive projects contain up to 1.6 million variables.

The smart recommendation system is developed to support the user during the process of adding hardware devices, i.e., creating new variables. For example, the system stores a template of device type IEDx99 (IED, intelligent electronic device) represented by the variables shown in Fig. 1. A user starts to add a new device IEDA19 by creating *two* new variables *zLD1/LLN0/Mod/stVal[ST]* and *zLD1/Q0CSWI1/Pos/Oper.AddCause* (Fig. 1, solid border) manually. The smart recommendation system will now find and suggest the remaining variables (Fig. 1, dashed border) by matching the two manually entered variables with the already known template of IEDx99. Suggesting and automatically creating the missing variables can save time and decrease the error rate during the projecting process.

[1]2nd International Data Science Conference 2019, Salzburg, Austria

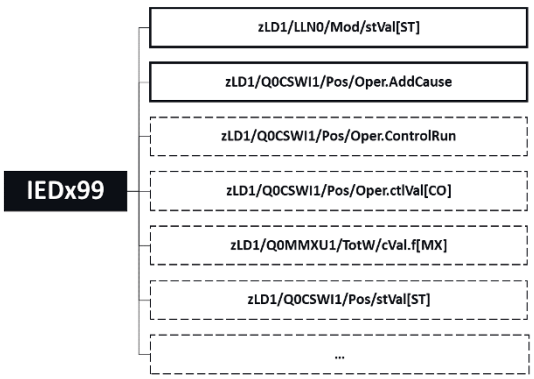


Fig. 1: Template of device IEDx99 – Variables with solid borders are created by the user; the remaining variables (dashed border) are recommended by the system.

II. Data Characteristics and Modelling Challenges

The data analysed in this study originates from real-world projects developed with *zenon* [1], a HMI/SCADA software platform used in several industries, e.g., automotive, energy, food and beverage. Tab. 1 shows an extract of the central variable list of a *zenon* project. These variables represent certain measured values or states of the hardware, including properties such as scaling or limit values. Beside the mandatory properties of a variable – a unique name, the driver enabling the communication with a device and the data type – optional metadata like measuring unit, number of decimals, limit values and many more can be set by the user.

In this study mainly data from energy industry projects are investigated. Projecting guidelines of *zenon* recommend applying industrial standards. The IEC61850 [3] standard developed for the power industry defines a convention applicable to the variable names. Exploring the data of different projects indicates that hardly anyone is implementing this naming convention; instead, each project seems to follow self defined rules for structured variable naming. By comparing the variable names of three different energy projects, it is illustrated that project A (see Tab. 1) follows

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_13

the standard while the names in project B (see Tab. 2) and project C (see Tab. 3) are structured differently. Using a well defined standard would enable an implicit classification, because the classes could be directly inferred from variable names.

Tab. 1: Example Data – Project A – List of variables in zenon HMI/SCADA software platform.

	Name	Drivers	DataType
0	...	...	...
1	IEDA18!zLD1/Q0CSWI1/Pos/Oper.ctlVal[CO]	IEC850	BOOL
2	IEDA18!zLD1/Q0CSWI1/Pos/stVal[ST]	IEC850	UDINT
3	IEDA18!zLD1/Q0MMXU1/TotW/cVal.f[MX]	IEC850	REAL_3d
4	**IEDA19!zLD1/LLN0/Mod/stVal[ST]**	**IEC850**	**SINT**
5	IEDA19!zLD1/Q0CSWI1/Pos/Oper.AddCause	IEC850	SINT
6	IEDA19!zLD1/Q0CSWI1/Pos/Oper.ControlRun	IEC850	SINT
7	IEDA19!zLD1/Q0CSWI1/Pos/Oper.ctlVal[CO]	IEC850	BOOL
8	IEDA19!zLD1/Q0CSWI1/Pos/stVal[ST]	IEC850	UDINT
9	IEDA19!zLD1/Q0MMXU1/TotW/cVal.f[MX]	IEC850	REAL_3d
10	**IEDB00!zLD1/LLN0/Mod/stVal[ST]**	**IEC850**	**SINT**
11	IEDB00!zLD1/Q0CSWI1/Pos/Oper.AddCause	IEC850	SINT
12	IEDB00!zLD1/Q0CSWI1/Pos/Oper.ControlRun	IEC850	SINT
13	IEDB00!zLD1/Q0CSWI1/Pos/Oper.ctlVal[CO]	IEC850	BOOL
14	IEDB00!zLD1/Q0CSWI1/Pos/stVal[ST]	IEC850	UDINT
15	IEDB00!zLD1/Q0MMXU1/TotW/cVal.f[MX]	IEC850	REAL_3d
16	**IEDB01!zLD1/LLN0/Mod/stVal[ST]**	**IEC850**	**SINT**
17	...	...	...

Since the data does not contain any labels it is not obvious which collection of variables represent a single hardware device. Sorting the variables by name and investigating the data manually reveals repeating patterns. For example, in Tab. 1 row 4 to 9 represent one device *IEDA19* with almost identical properties – data type, measuring unit, limit values, etc. as well as large parts of the name – to device *IEDB00* from row 10 to 15.

Tab. 2: Example Data – Project B – List of variables in zenon HMI/SCADA software platform.

	Name	Drivers	DataType
...	...	...	...
0	110KV_DIR001_BUS01_CO	IEC850	USINT
1	110KV_DIR001_BUS01_RV	IEC850	USINT
2	110KV_DIR001_BUS02_CO	IEC850	USINT
...	...	...	...
15	110KV_DIR001_MX_U	IEC850	REAL
16	110KV_DIR001_XCBR_CO	IEC850	USINT
17	110KV_DIR001_XCBR_RV	IEC850	USINT
...	...	...	...

The first intention is using unsupervised learning methods to create templates of the devices for the classification as

Tab. 3: Extracted Data – Project C – List of variables in zenon HMI/SCADA software platform.

	Name	Drivers	DataType
...	...	...	...
0	E.SCA.20kV.CCA..Anregung Erde	IEC850	BOOL
1	E.SCA.20kV.CCA..Anregung L1	IEC850	BOOL
2	E.SCA.20kV.CCA..Anregung L2	IEC850	USINT
...	...	...	...
58	E.SCA.20kV.CCA..U/AMZ Generalauslösung	IEC850	BOOL
59	E.SCA.20kV.CCA..Warnungssammelmeld	IEC850	BOOL
60	E.SCA.20kV.CCA..Wirkleistung	IEC850	REAL
...	...	...	...

mentioned in section I. An exploratory data analysis (EDA) indicates that the variables of a hardware device (e.g., IEDA19) do not have common properties; therefore, no similarities can be found. Even the variable names contain only short identical chunks. Assigned to the example data of Tab. 1 a cluster *c1* would contain the variables in row 4, 10 and 16, cluster *c2* would contain 5, 11, 17 and so forth. These groups have more in common – large parts of the name and other variable properties – than the variables belonging to one hardware device. Due to its nature, that is grouping objects with similar characteristics [7, p. 118], clustering does not perform very well here. Thus, for further investigations it is assumed that predefined templates representing classes of devices are already given and the supervised classification problem is addressed:

- Is it possible to preprocess data appropriately such that supervised methods can be applied at all?
- Is it possible to classify a bunch of new variables correctly and predict the remaining ones?
- Which features provide sufficient information to solve the classification problem?

III. Model Assumptions

A previous thesis [9], investigating the same topic, proposes a solution using string metrics to classify devices. The Sørensen–Dice coefficient [2], [8] is used to find similarity in variable names and the Hungarian algorithm [5] to find the appropriate predefined device template matching the variables. Results show that the successful application of the recommendation system highly depends on a valid and consistent variable naming convention.

To compensate for the weakness of having a naming convention, in this study, the classification should be based on variable properties only, e.g., measuring unit, number of decimals, limit values, addressing information. Related variable names should not be considered as features. Furthermore, the data set consisting of qualitative and quantitate features needs preprocessing such that supervised methods can be applied. For example, row 4 to 9 in Tab 1 define one device – the data needs to be transformed such that there is one row per

hardware device. We assume that the qualitative data contains more information than the quantitative one. Therefore, we propose a model aggregation where both data sets are first modelled independently, but finally combined. Two different approaches are investigated using (1) one-hot encoding [4] and (2) probabilistic topic models [6]. Using the first approach, the feature matrix is generated using one-hot encoding. Tab. 4 shows the resulting matrix for one property, namely data type. The number of one-hot encoded features is the sum of the cardinality of all properties. As there is not only one property considered for the one-hot encoding, but approximately 40, this might imply a sparsity problem. New variables are then classified by having a hamming distance equal to zero.

Tab. 4: One-hot encoding for the property data type of IEDx99 template.

	SINT	UDINT	BOOL	Real_3d
...!zLD1/LLN0/Mod/stVal[ST]	1	0	0	0
.../Q0CSWI1/Pos/Oper.AddCause	1	0	0	0
.../Q0CSWI1/Pos/Oper.ControlRun	1	0	0	0
.../Q0CSWI1/Pos/Oper.ctlVal[CO]	0	0	1	0
...!zLD1/Q0CSWI1/Pos/stVal[ST]	0	1	0	0
.../Q0MMXU1/TotW/cVal.f[MX]	0	0	0	1

For the probabilistic topic modelling approach it is necessary to transform the data. Tab. 5 shows a generic matrix for a device IED_1 consisting of one row per variable and one column per property (e.g., data type, measuring unit, limit values). To apply topic modelling the matrix has to be vectorized to $v(IED_1)$ (see equation 1). As there are several devices in the training set, vectorization is applied for each device (class), i.e., every device is represented by a unique vector of words.

Tab. 5: Generic matrix for one device IED_1 will be vectorized to $v(IED_1)$.

	$\mathbf{Prop_1}$	$\mathbf{Prop_2}$	$\mathbf{Prop_3}$	...	$\mathbf{Prop_k}$
$\mathbf{Var_1}$	prop_{11}	prop_{12}	prop_{13}	...	prop_{1k}
$\mathbf{Var_2}$	prop_{21}	prop_{22}	prop_{23}	...	prop_{2k}
...	...	...	...	...	...
...	...	...	...	...	...
$\mathbf{Var_m}$	prop_{m1}	prop_{m2}	prop_{m3}	...	prop_{mk}

$$v(IED_1) = (\text{prop}_{11}, ..., \text{prop}_{1k}, \text{prop}_{21}, ..., \text{prop}_{2k}, ..., \text{prop}_{mk}) \quad (1)$$

For the classification of a bunch of m new variables, the variables are also vectorized, resulting in $v(IED') = (\text{prop}'_{11}, ..., \text{prop}'_{1k}, \text{prop}'_{21}, ..., \text{prop}'_{mk})$. Recommended classes are then devices, where the vectors have a matching cardinality of $m \cdot k$. Adding probabilistic to the methodology leads to a ranking of recommended devices and might compensate for inconsistencies within same device types, e.g., different versions with deviations in associated properties.

IV. Summary

In this paper the problem of developing a smart recommendation system to support the projecting process in HMI/SCADA software platforms is investigated. Data characteristics of real-world projects and modelling challenges are examined. Due to the nature of the data, it is unlikely to cluster variables representing a device properly. Thus, without a consistent naming of the variables manual labelling of the data is inevitable. Furthermore, two supervised learning methods are proposed, i.e., one-hot encoding and topic modelling to solve the problem. The results and performance of these models will be presented in future studies.

Acknowledgment: This work was partly funded by the Austrian Research Promotion Agency FFG, Project No. 860946.

References

[1] *zenon Software Platform*, https://www.copadata.com/hmi-scada, Accessed: 08.04.2019.
[2] Lee R. Dice, *Measures of the Amount of Ecologic Association Between Species*, 1945.
[3] *Communication networks and systems for power utility automation – Part 7-1: Basic communication structure – Principles and models*, Standard, International Electrotechnical Commission, July 2011.
[4] ChinmayD.Pai K. Potdar, TaherS.Pardawala, *A comparative study of categorical variable encoding techniques for neural network classifiers*, International Journal of Computer Applications **175** (2017), no. 4.
[5] H. W. Kuhn, *The Hungarian method for the assignment problem*, 1955.
[6] T. Griffiths M. Steyvers, *Probabilistic topic models. Handbook of latent semantic analysis*, 2007.
[7] EMC Education Services (ed.), *Data science and big data analytics: Discovering, analyzing, visualizing and presenting data*, Wiley, 2015.
[8] T. Sørensen, *A Method of establishing groups of equal amplitude in plant sociology based on similarity of species content*, 1948.
[9] M. Thomas, *Intelligente Vorschlagsfunktion als Unterstützung für die Entwicklung von Projekten eines HMI/SCADA-Softwaresystems*, Master's thesis, Salzburg University of Applied Sciences, 8 2018.

Adversarial Networks — A Technology for Image Augmentation

Maximilian Ernst Tschuchnig
Salzburg University of Applied Sciences
Puch bei Hallein, Austria
maximilian.tschuchnig@fh-salzburg.ac.at

***Abstract*—A key application of data augmentation is to boost state-of-the-art machine learning for completion of missing values and to generate more data from a given dataset. In addition to transformations or patch extraction as augmentation methods, adversarial networks can be used to learn the probability density function of the original data. Generative adversarial networks (GANs) are an adversarial method to generate new data from noise by pitting a generator against a discriminator and training in a zero-sum game trying to find a Nash Equilibrium. This generator can then be used in order to convert noise into augmentations of the original data. This short paper shows the usage of GANs in order to generate fake face images as well as tips to overcome the notoriously hard training of GANs.**

***Index Terms*—Data Augmentation, Adversarial Networks, Face Image Generation, GAN Training**

I. Introduction

Data augmentation is a common method to boost state-of-the art machine learning approaches, especially if the amount of training data is relatively small. Most of the time, this is accomplished by increasing the number of samples from the given, relatively small, ground truth through basic transformations like translation, rotation [1] or patch extraction [2]. Data augmentation can be used to increase machine learning capabilities in applications where representative data is hard to obtain or sufficient amounts of data have not been collected already. One such application might be in industrial machine learning applications like predictive maintenance, which a lot of times is needed in industrial applications without data being captured or captured in an unusable form. Since machine learning uses a lot of data, which is often non-existent or in an unusable format, augmenting usable data is of importance. In addition to conventional data augmentation, adversarial networks, consisting of a discriminator D and a generator G, can be used in order to generate data.

In this short paper, the author gives an introduction into generative adversarial networks (GANs), shows a demo implementation (using Python and Keras) of generating human faces and gives tips in order to ease the training of GANs.

II. Method and Implementation

Fig 1 illustrates a basic deep convolutional GAN (DCGAN) which uses Neural Networks and Convolutions in D as well as G. The figure also shows the basic structure of a GAN, which consists of G trying to fool D into believing that the data generated by G from a random noise z is part of the real data $x \in X$. G is successful if $D(G(z)) = Real$, while D is successful if $D(G(z)) = Fake$ and $D(x) = True$. The simultaneous training of D and G to minimize the log likelihood from the functions above, leads to a two player minmax game [3] with the Kullbach Leibler (KL) Divergence as the value function. GANs, as proposed by [4], typically use a symmetric version of the KL divergence, the Jensen Shannon divergence.

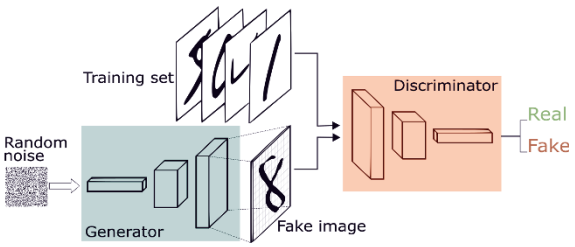


Fig. 1. GAN structure [5]

The data that the author aims to augment are celebrity faces from the online accessible celeb faces [6] dataset. This data was chosen as a demonstration, a real world application of this data augmentation technique would be to generate heatmaps from one dimensional sensor data with subsequent classification.

The given link[1] shows the structure of the celebrity faces generator as well as the discriminator, which is a basic two class convolutional neural network using 2D convolutions, batch norm, leaky ReLUs and a fully connected layer. The generator creates 64x64 data from 8x8 random noise. It uses transposed convolution [7] for upscaling, as well as batch norm and leaky ReLUs. Alternatively to transposed convolutions, 2D up sampling with a following convolution can also be used for a similar effect. Also note how the proposed generator does not make use of pooling or dropout [4]. Listing 1 shows the final composition of D and G in order to recreate the structure given in Fig 1. In order to update D and G on their own, both networks are created and the discriminator weights are frozen before adding it into a keras sequential. This enables us to train D directly and G by inserting noise to the GAN without updating D.

[1] https://github.com/kuchenkiller/FaceGenerator/blob/master/Appendix.pdf

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_14

Code Listing 1. GAN Composition

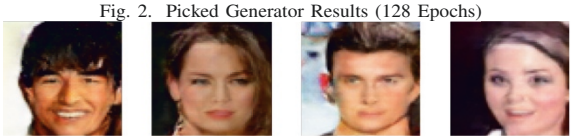
```
discriminator = make_discriminator(leaky, std)
discriminator.compile(optimizer=Adam(lr=d_learn,
        beta_1=d_beta), loss='binary_crossentropy')
make_trainable(discriminator, False)

generator = make_generator(sample_size, leaky, std)

gan = Sequential([generator, discriminator])
gan.compile(optimizer=Adam(lr=g_learn,
        beta_1=g_beta), loss='binary_crossentropy')
```

III. Discussion and Training Tips

Fig. 2. Picked Generator Results (128 Epochs)



Using this architecture, the results of the original implementation [6] can be reproduced with basic hardware[2], as the example images in Figure 2 illustrate. A further experiment with the MNIST dataset shows that the method generalizes to other tasks. However, since the hyperparameter setting is non-trivial, the following lessons learned should be considered to improve the learning outcomes and additionally were tested on heatmaps constructed from household power usage:

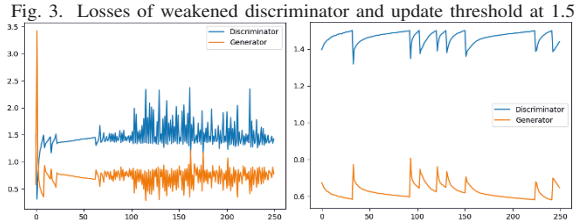
Dominating Discriminator — While training GANs, a dominating D, resulting in useless gradients, seems to be common problem. To achieve such a behaviour the amount of convolutions in the dominating network can be reduced, effectively crippling D. Furthermore, the learning rate can be decreased and the learning rate decay increased [8]. It should be noted that the best optimization results were observed by finding a well constructed network architecture, mostly by crippling D. Both the celeb faces as well as heatmap picture generating GANs were able to be parametrised successfully by applying this knowledge (Fig 3 shows an example of this).

Dominating Networks in General — Other ways to counter dominating networks are to add noise to real image labels (for example $noise = 0.1 \rightarrow true \neq 1$ but $0.9 < true < 1$) and to set a minimal loss threshold which has to be met in order to update the dominating network weights. Adding noise to true labels works twofold. Adding noise prevents $[0, 1]$ trivial solutions, leading the KL divergence to be finite. Also, the generator can never perfectly describe the real data with the label 1. Adding noise makes generated data more feasible to the discriminator [9]. Setting a min update threshold is similar to separate update epochs, but in a dynamic way. The dominating part is only trained if the previous train or test step (depending on the length of an epoch) exceeds a certain threshold. Fig 3 shows the celeb faces as an example which tend to have a dominating discriminator and epochs tend to take a long time. In order to implement such an update threshold one can evaluate the current epoch step and only apply the actual training (train on batch) if the evaluation loss is above a certain threshold.

Multiple Separate Update Epochs — Another way to improve GAN training is to train D and G in separate epochs in order to not continue working into local minima. This can be accomplished by alternating train on batch between D and G on whole epochs. Since epoch lengths influence the usability of this method, they have to be taken into account. While applying this method makes a lot of sense if several epochs are calculated very quickly (tested examples: MNIST and Heatmaps) it was counterproductive on very few, long epochs (celeb faces). This behaviour can also be used on iterations instead of epochs in the case of long epochs.

Early stopping — Especially when running an exhaustive grid search or in general when run time is an issue, early stopping should be implemented. Since GAN losses are different to conventional deep learning losses (lower losses do not represent a good approximation of the problem), a different kind of early stopping had to be applied. The implementation chosen by the author was to keep track of the last n epochs losses and calculate the mean from these values. The training procedure is stopped if the losses mean stays between a threshold T_L for n epochs.

Fig. 3. Losses of weakened discriminator and update threshold at 1.5



[2]Hardware: 12 core i7 3,2GHz and two GeForce GTX 980

References

[1] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *arXiv preprint arXiv:1712.04621*, 2017.

[2] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, and H. H. Saltz, "Patch-based convolutional neural network for whole slide tissue image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2424–2433.

[3] J. F. Nash, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. l. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[5] S. T., "An intuitive introduction to generative adversarial networks (gans)," https://medium.freecodecamp.org/an-intuitive-introduction-to-generative-adversarial-networks-gans, 2018, accessed: 2019-04-05.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[7] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, pp. 19–24, 2016.

[8] M. Arjovsky and L. Bottou, "Towards principled methods for training generative adversarial networks," *arXiv preprint arXiv:1701.04862*, 2017.

[9] C. K. Sønderby, J. Caballero, L. Theis, W. Shi, and F. Huszár, "Amortised map inference for image super-resolution," *arXiv preprint arXiv:1610.04490*, 2016.

Using supervised learning to predict the reliability of a welding process

Melanie Zumtobel
FH Vorarlberg
Dornbirn, Austria
melanie.zumtobel@fhv.at

Kathrin Plankensteiner
FH Vorarlberg
Dornbirn, Austria

***Abstract*—In this paper, supervised learning is used to predict the reliability of manufacturing processes in industrial settings. As an example case, lifetime data has been collected from a special device made of sheet metal. It is known, that a welding procedure is the critical step during production. To test the quality of the welded area, End-of-Life tests have been performed on each of the devices.**

For the statistical analysis, not only the acquired lifetime, but also data specifying the device before and after the welding process as well as measured curves from the welding step itself, e.g., current over time, are available.

Typically, the Weibull and log-normal distributions are used to model lifetime. Also in our case, both are considered as an appropriate candidate distribution. Although both distributions might fit the data well, the log-normal distribution is selected because the ks-test and the Bayesian Factor indicate slightly better results.

To model the lifetime depending on the welding parameters, a multivariable linear regression model is used. To find the significant covariates, a mix of forward selection and backward elimination is utilized. The t-test is used to determine each covariate's importance while the adjusted coefficient of determination is used as a global Goodness-of-Fit criterion. After the model that provides the best fit has been determined, predictive power is evaluated with a non-exhaustive cross-validation and sum of squared errors.

The results show that the lifetime can be predicted based on the welding settings. For lifetime prediction, the model yields accurate results when interpolation is used. However, an extrapolation beyond the range of available data shows the limits of a purely data-driven model.

I. Introduction

In reliability analysis, End-of-Life (EoL) tests are necessary to guarantee that products operate reliably. Since it is not possible to test the devices at real stress conditions, accelerated stress tests in combination with statistical models are commonly applied to achieve reliable forecasts for the lifetime of the devices. This methodology of reliability analysis is already well-established and commonly applied in fields with high safety demands like aerospace, nuclear power generation but also automobiles, computers, and cell phones. Nevertheless, also for markets with lower safety demands, there is a need to have high reliability at minimal cost. Companies that are able to provide highly reliable products without excessive cost will be the most successful ones. Finding a good balance requires appropriate reliability analyses and related decision making [4], [5].

In this study, reliability analysis in combination with statistical modeling is used to gather information about a special device made of sheet metal. With this, the optimal setting for welding in a mass production is defined. In combination with real-time process data, the model which was developed will enable inference about the quality of the product during manufacturing. Devices with an insufficient quality will be detected as soon as possible and expensive resources for final assembly will be saved.

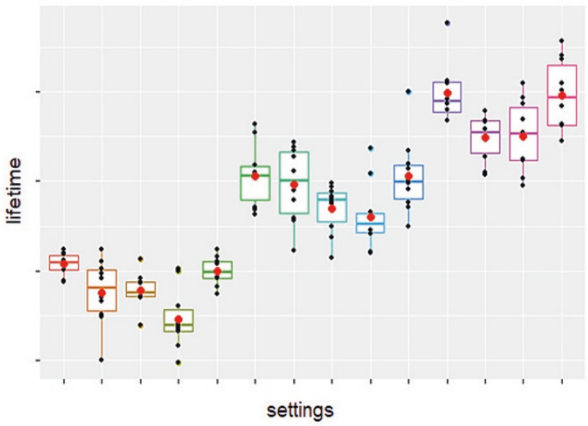


Fig. 1: Lifetime distributions of 14 welding settings.

II. Data Characteristics

The lifetime data comes from one device type, welded with different combinations of current, welding time, and bonding force. The variation of these parameters result in different power and energy curves during welding. The combinations of welding settings were determined by an a-priori study and a Design-of-Experiment (DoE) [6]. After the welding process, the devices are tested with accelerated stress tests until EoL. Lifetime is thereby measured in cycles to failure (CTF). Based on the variation of test settings, different lifetime distributions are observed, see Fig. 1.

Based on the fact that Weibull and log-normal distribution are typically used when modeling lifetime data, both where considered for this analysis. From a pure visual point of view,

P. Haber et al. (Hrsg.), *Data Science – Analytics and Applications*,
https://doi.org/10.1007/978-3-658-27495-5_15

Weibull and log-normal distribution fit empirical data quite well: Fig. 2 shows empirical against theoretical quantiles of both distributions for one welding setting. It indicates a good fit for both candidates. To determine which distribution should be used for modeling, not only statistical methods but also physical factors and the current understanding of the welding process are considered [2], [7].

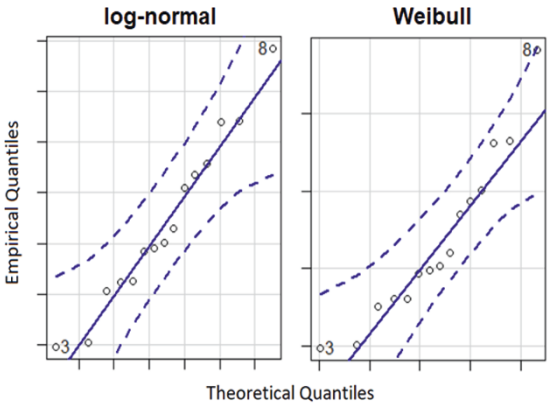


Fig. 2: Empirical and theoretical quantiles assuming a log-normal (left) and Weibull distribution (right). One welding setting is shown.

From a statistical point of view, p-values obtained by a Kolmogorov-Smirnov test (ks-test) and Bayes Factors (BF) [3] don't give a consistent evidence which distribution should be taken. Table 1 gives an overview of skewness (γ), kurtosis (κ), p-values of a ks-test (p_{logN} and p_{Weib}) and BF for each lifetime distribution. It shows that empirical data can be left ($\gamma < 0$) or right ($\gamma > 0$) skewed and that the observed distributions can be platykurtic ($\kappa < 3$) or leptokurtic ($\kappa > 3$). Furthermore, the p-values (p_{logN} and p_{Weib}) of the ks-test indicate, that none of the considered distributions can be rejected. BF give evidence that at least for setting 8 (see Table 1) a log-normal distribution fits the empirical data better than a Weibull distribution (BF > 3). For all other settings, the Goodness-of-Fit of both candidate distributions is comparable, indicated by $1/3 <$ BF < 3.

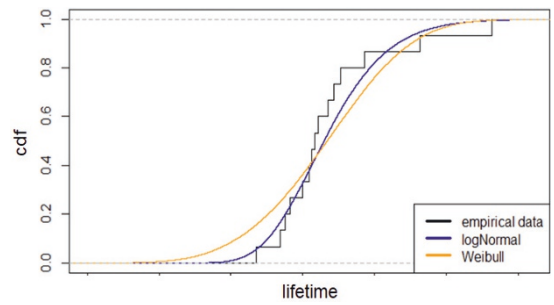


Fig. 3: Cumulative distribution function (cdf) of empirical data with fitted log-normal and Weibull distribution. One welding setting is shown.

Fig. 3 shows the cumulative distribution function (cdf) of both candidates for one welding setting - both candidates fit the empirical data well.

Tab. 1: Overview of skewness, kurtosis, p-values of a ks-test considering log-normal (p_{logN}) and Weibull (p_{Weib}) as well as the Bayes Factor (BF) of each lifetime distribution. Each lifetime distribution represents one welding setting.

setting	skewness	kurtosis	pLogN	pWeib	BF
1	-0.24	1.92	0.98	0.96	1.45
2	-0.32	2.32	0.92	0.93	1.85
3	0.13	3.54	0.54	0.67	0.62
4	0.85	2.59	0.59	0.42	0.59
5	0.06	2.16	1.00	0.98	0.93
6	0.60	2.15	0.92	0.78	0.79
7	-0.18	1.62	0.86	0.89	1.95
8	-0.80	2.38	0.36	0.87	4.25
9	1.28	3.55	0.56	0.37	0.44
10	1.59	5.02	0.88	0.66	0.36
11	1.65	4.60	0.69	0.48	0.49
12	-0.35	1.78	0.95	0.96	2.09
13	0.24	1.82	0.95	0.89	1.05
14	0.44	1.79	0.75	0.78	0.89

From a physical point of view, a multiplicative effect behavior is more likely than an additive one. Thus, modeling with a log-normal distribution might be more reasonable [7].

Therefore, to describe and predict the lifetime of the welded devices, a log-normal distribution is selected, empirical data is log-transformed and the $logCTF$ is modeled.

III. Model Development and Evaluation

To find an appropriate model that can describe the data efficiently and accurately, supervised learning, particularly a multivariable linear regression model, is used [1]:

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_j x_{ij} + ... + \beta_J x_{iJ} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

where

$$\begin{aligned} Y_i &= logCTF \text{ for } i\text{-th test setting (response)} \\ \beta_0 &= \text{intercept} \\ \beta_j &= \text{coefficient of } j\text{-th feature (covariate)} \\ x_{ij} &= j\text{-th feature (covariate) of observation } i. \end{aligned}$$

A. Model Selection & Model Evaluation

For model selection, two different Goodness-of-Fit criteria are used. For the local Goodness-of-Fit, the t-test is applied. The resulting p-value is used to decide if the associated covariate is important for the explorative power of the model. Small p-values indicate, that the covariate is significant and should be used to model the lifetime $logCTF$ [1]. As a global Goodness-of-Fit criterion, in this study the adjusted coefficient

of determination (R^2_{adj}) is used [8]. Alternative Goodness-of-Fit criteria such as standardized sum of squared errors or F-statistic give similar results. To find the significant features (covariate selection), a combination of forward selection and backward elimination (= hybrid approach) is utilized [8]. In every step, a covariate is added to the model - one covariate at a time. The model is trained and the significance of each covariate (local Goodness-of-Fit) is evaluated. Once a covariate loses its importance (indicated by a p-value > 0.5), it is neglected for the next step. This procedure is repeated until the defined global Goodness-of-Fit criterion reaches an optimum – meaning that the best model has been found.

Using this procedure, the lifetime can be modeled by

$$E[logCTF] = \beta_0 + \beta_1 * I + \beta_1 * E,$$

where I is the maximum current and E the maximum energy during the welding process.

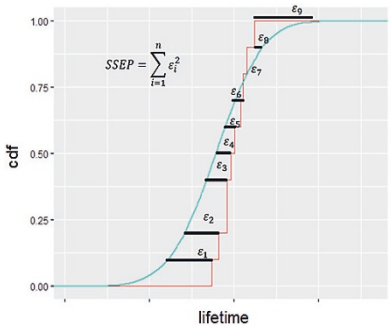


Fig. 4: Quantile associated errors and sum of squared errors of prediction (SSEP).

After the model with the best fit has been selected, k-fold cross-validation [1] has been applied and errors between observed and estimated lifetime (regarding the quantiles) are calculated [1]. Fig. 4 shows the determination of errors and the sum of squared errors of prediction (SSEP) for one welding setting. The red and blue curve denote the cdf of empirical and theoretical data, respectively. Although the model lacks accuracy in some rare cases, it shows promising results indicated by an error distribution following a standard normal distribution overall (see Fig. 5).

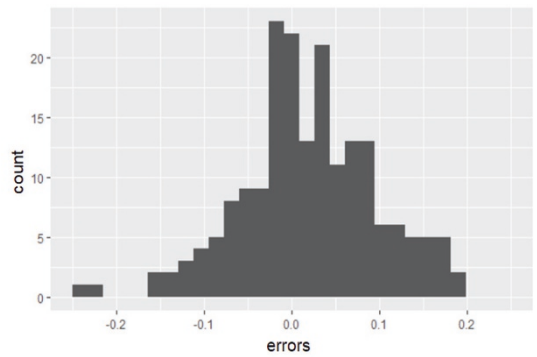


Fig. 5: Error distribution of cross-validation.

B. Prediction of Lifetime according to new Settings

Based on the fact, that there were still some welded devices left for lifetime testing, the regression model has been used to forecast those missing data points. For this purpose, a regression model based on maximum current and maximum energy has been trained with all observed data. The trained model has then been used for prediction.

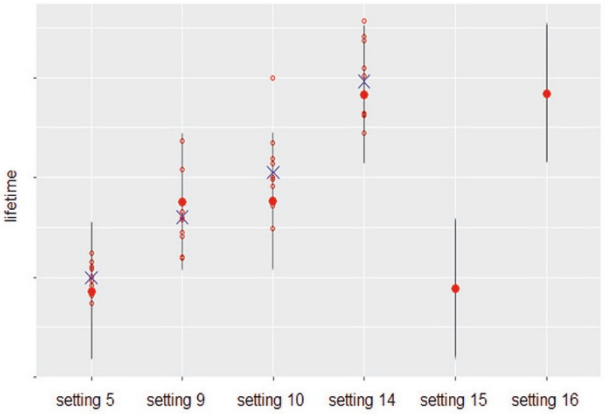


Fig. 6: Prediction results for old and new settings.

Fig. 6 shows the predictions for settings 5, 9, 10, and 14 as well as for two completely new welding settings (setting 15 and setting 16). The red dots and black line represent the predicted mean lifetime and 95% prediction interval, respectively. The red circles denote already observed data for the corresponding setting, the blue crosses denote their mean. It can be seen that for setting 5, 9, 10, and 14 lifetime data has been already available and used for the training. In these cases, it is not astonishingly that the predictive power of the model is high. More interesting is the predictive performance for setting 15 and 16. Fig. 7 and Fig. 8 show the empirical and predicted cdf for these settings.

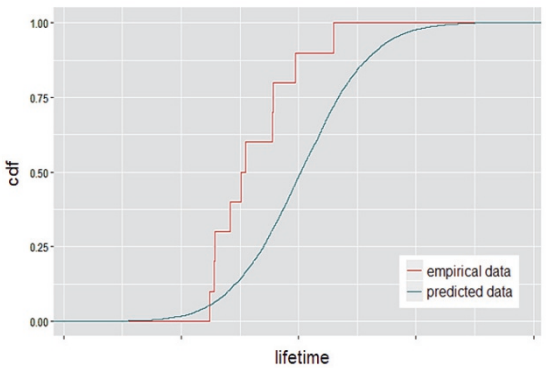


Fig. 7: Cumulative distribution function (cdf) of empirical and predicted lifetime for setting 15.

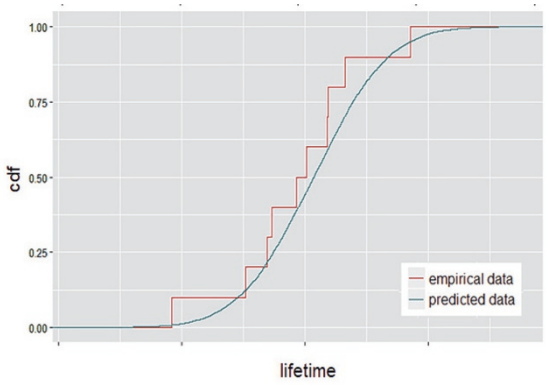


Fig. 8: Cumulative distribution function (cdf) of empirical and predicted lifetime for setting 16.

While the predictive power for setting 16 is high, prediction for setting 15 lacks accuracy. One reason for this could be that a purely data-driven model cannot reflect the physics and therefore, it commonly lacks accuracy for extrapolative tasks. Combining subject-matter knowledge (e.g., about the physics of failure) might compensate for this weakness [4].

IV. Summary

In this paper, multivariable linear regression has been used to model and predict the reliability of a welding process. For this purpose, welded devices have been tested with accelerated stress tests to obtain associated lifetime data.

To model the lifetime, the log-normal and Weibull distribution have been investigated. Although both distributions fit the data well, the log-normal distribution has been selected, since it gives slightly better results. To find the significant covariates that influence the mean lifetime, a mix of forward selection and backward elimination has been utilized. Thereby, the t-test has been used to determine each covariate's importance (local Goodness-of-Fit) while the adjusted coefficient of determination has been used to determine the significance of the whole regression model (global Goodness-of-Fit).

With that, it could be determined that the mean lifetime can be modeled depending on current and energy. The model's predictive power has been evaluated with a non-exhaustive cross-validation and SSEP. Using that model, further results of tests could be forecasted accurately. However, an extrapolation beyond the range of available data shows the limits of a purely data-driven model. Combining subject-matter knowledge (e.g., about the physics of failure) could compensate for this weakness.

This study enabled to find the optimal welding setting for the upcoming mass production. In the future, this pre-trained model will be used to evaluate and predict the quality of the welding process in real-time. With time (and appropriate data), the model will get more and more accurate and costly test resources can be saved without compromising quality and reliability.

References

[1] Klaus Backhaus, Bernd Arichson, Wulff Plinke, and Rolf Weiber, *Multivariate Analysemethoden*, vol. 14, Springer Gabler, 2016.

[2] Arabin Kumar Dey and Debasis Kundu, *Discriminating Among the Log-Normal, Weibull, and Generalized Exponential Distributions*, IEEE Transactions on Reliability, vol. 58(3), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.562.2820&rep=rep1&type=pdf, 2009.

[3] Leonhard Held, *Methoden der statistischen Inferenz: Likelihood und Bayes*, Springer Spektrum, 2008.

[4] William Q. Meeker, *Bill Meeker on reliability in the age of big data*, https://community.jmp.com/t5/JMP-Blog/Bill-Meeker-on-reliability-in-the-age-of-big-data/ba-p/62497, 2018.

[5] William Q. Meeker and Yili Hong, *Reliability Meets Big Data: Opportunities and Challenges*, Journal of Quality Engineering, https://lib.dr.iastate.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1085&context=stat_las_preprints, 2014.

[6] Douglas C. Montgomery, *Design and analysis of experiments*, vol. 8, John Wiley and Sons, 2013.

[7] Jesus Francisco Ortiz-Yanez and Manuel Roman Pina-Monarrez, *Discrimination between the lognormal and Weibull distributions by using multiple linear regression*, DYNA, 85(205), http://www.scielo.org.co/pdf/dyna/v85n205/0012-7353-dyna-85-205-00009.pdf, 2018.

[8] Petra Stein, Monika Pavetic, and Marcel Noack, *Multivariate Analyseverfahren*, https://www.uni-due.de/imperia/md/content/soziologie/stein/multivariate.pdf, 2011.